MICROSATELLITE EVOLUTION IN THE YEAST GENOME - A GENOMIC APPROACH

A thesis submitted in fulfilment of the requirements for the Degree

of

Doctor of Philosophy

in Biology

by

Angelika Merkel

University of Canterbury

2008

“If we have learned anything at all in a century and a half of evolutionary biology, that facile generalizations are dangerous. The evolutionary process finds a way to create exceptions to every model we propose”.

A.L. Hughes (2005)

Acknowledgements

First of all I would like to thank my supervisor Prof. Neil J. Gemmell for giving me the opportunity to pursue this PhD and for all the other opportunities that allowed me to further develop my skills and myself as a scientist. I appreciated his human attitude and patience, his talent for multitasking, and particular, his incredible ability to work under extreme time pressure. Thanks also to Prof. Jack Heineman for refreshing conversations during the early days of my PhD, and Dominic Lee, Dave Kelly, Jason Tylianakis, and Raazesh Sainudiin for statistical advice and other insights during the later ones. Thanks to Joel Pitt and Nicola Day for help with the initial R-scripts, and Andrew Bagshaw for help with the wavelet analysis.

Special thanks go to Matt Walters for help with figures, poster and always making time to get stuff done, and to Ronny Groenteman and Michael Sarfati for R trouble shooting and cheer-ups. Many thanks also to Winston Hyde and everyone from the programming course at Cold Spring Harbor Laboratory in 2007 for keeping the spirit alive.

The Molecular Ecology Lab has been a great place to work, a credit to all those wonderful characters it was made of. Best wishes for all of you and your future careers. I am looking forward to meeting again sometime, somewhere…

I am indebted to my friends that have not given up on me and endured me even during some of the darker moments. Thanks Sarah, thanks Liz, thanks to all 'the folks at No. 28'. To my family, especially mum and Grandma Ruth, I am grateful - thanks for believing in me, it meant a lot.

Finally, my partner, David John Hood, the 'rock' in my life. Thanks for bearing, loving, caring, the lunch packages you made and all those chocolates you brought me. It would have been a lot harder without you.

Table of Contents

List of Tables

List of Figures

Abbreviations

ARS	Autonomous Replicating Sequence
CDS	Coding Sequence
DBS	Double-Strand Break
GLM	Generalized Linear Model
Kb	Kilobase pair
LTR	Long Terminal Repeat
Mb	Megabase pair
MMR	Mismatch Repair
ORF	Open Reading Frame
QTL	Quantitative Trait Loci
SGD	Saccharomyces Genome Database
SGRP	Saccharomyces Genome Resequencing Project
SNP	Single Nucleotide Polymorphism
TRF	TandemRepeatFinder
UTR	Untranslated Region

Abstract

Microsatellites are short (1-6bp long) highly polymorphic tandem repeats, found in all genomes analyzed so far. Popular genetic markers for many applications including population genetics, pedigree analysis, genetic mapping and linkage analysis, some microsatellites also can cause a variety of human neurodegenerative diseases and may act as agents of adaptive evolution through the regulation of gene expression. As a consequence of these diverse uses and functions, the mutational and evolutionary dynamics of microsatellite sequences have gained much attention in recent years. Mostly, the focus of studies investigating microsatellite evolution has been to develop more refined evolutionary models for estimating parameters such as genetic distance or linkage disequilibrium. However, there is an incentive in using our understanding of the evolutionary processes that affect these sequences to examine the functional implications of microsatellite evolution. What has emerged from nearly two decades of study are highly complex mutational dynamics, with mutation rates varying across species, loci and alleles, and a multitude of potential influences on these rates, most of which are not yet fully understood.

The increasing availability of whole genome sequences has immensely extended the scope for studying microsatellite evolution. For example, where once it was common to examine single loci, it is now possible to examine microsatellites using genome wide approaches. In the first part of my dissertation I discuss approaches and issues associated with detecting microsatellites in genomic data. In **Chapter 2** I undertook a meta-analysis of studies investigating the distribution of microsatellites in yeast and showed that studies comparing the distribution of microsatellites in genomic data can be fraught due to the application of different definitions for microsatellites by different investigators. In particular, I found that variation in how investigators choose the repeat unit size of a microsatellite, handle imperfections in the array and especially the choice of minimum array length used, leads to a large divergence in results and can distort the conclusions drawn from such studies, particularly where inter-specific comparisons are being made. In a review of the currently available suite of bioinformatics tools (**Chapter 3**), I further

showed that this bias extends beyond a solely theoretical controversy into a methodological issue because most software tools not only incorporate different definitions for the key parameters used to define microsatellites, but also employ different strategies to search and filter for microsatellites in genomic data. In this chapter I provide an overview of the available tools and a practical guide to help other researchers choose the appropriate tool for their research purpose.

In the second part of my thesis, I use the analytical framework developed from the previous chapters to explore the biological significance of microsatellites exploiting the well annotated genome of the model organism *Saccharomyces cerevisiae* (baker's yeast). Several studies in different organisms have indicated spatial associations between microsatellites and individual genomic features, such as transposable elements, recombinational hotspots, GC-content or local substitution rate. In **Chapter 4**, I summarized these studies and tested some of the underlying hypotheses on microsatellite distribution in the yeast genome using Generalized Linear Models (GLM) and wavelet transformation. I found that microsatellite type and distribution within the genome is strongly governed by local sequence composition and negative selection in coding regions, and that microsatellite frequency is inversely correlated with SNP density reflecting the stabilizing effect point mutations have on microsatellites. Microsatellites may also be markers for recent genome modifications, due to their depletion in regions nearby LTR transposons, and elements of potential structural importance, since I found associations with features such as meiotic double strand breaks, regulatory sites and nucleosomes. Microsatellites are subject to local genomic influences, particularly on small (1-2kb) scales. Although, these local scale influences might not be as dominant as other factors on a genome-wide scale they are certainly of importance with respect to individual loci.

Analysis of locus conservation across 40 related yeast strains (**Chapter 5**) showed no bias in the type of microsatellites conserved, only a negative influence of coding sequences, which supports again the idea that microsatellites evolve neutrally. Polymorphism was rare, and despite a positive correlation with array length, there was no

relationship with either genomic fraction or repeat size. However, the analysis also revealed a non-random distribution of microsatellites in genes of functionally distinct groups. For example, conserved microsatellites (similar to general microsatellites in yeast) are mostly found in genes associated with the regulation of biological and cellular processes. Polymorphic loci show further an association with the organization and biogenesis of cellular components, morphogenesis, development of anatomical structures and pheromone response, which, is absent for monomorphic loci. Whether this distribution is an indication of functionality or simply neutral mutation (e.g. genetic hitch-hiking) is debatable since most conserved microsatellites, particularly variable loci, are located within genes that show low selective constraints. Overall, microsatellites appear as neutrally evolving sequences, but owing to the sheer number of loci within a single genome, individual loci may well acquire some functionality. More work is definitely needed in this area, particularly experimental studies, such as reporter-gene expression assays, to confirm phenotypic effects.

Chapter 1

General Introduction

1.1. Microsatellites

Microsatellites are short (1-6bp) tandemly repeated DNA sequences, highly polymorphic and ubiquitously abundant in most genomes sequenced so far. Together with the longer (>10bp motif length) minisatellites they also have been termed as variable number tandem repeats (VNTRs). However, whereas minisatellites commonly evolve through recombination (Jeffreys *et al.* 1994), microsatellite instability is mainly caused by strand slippage during replication (Levinson and Gutman 1987b). The extraordinary high mutation rate (10^{-2} to 10^{-6} mutations per locus per generation) of microsatellites has made these simple sequence repeats (SSR) popular genetic markers for many applications including genetic mapping, population genetics, phylogenetics, pedigree analysis and even DNA forensics (Goldstein and Schlotterer 1999). Moreover, microsatellites have been shown to cause at least 30 human neurodegenerative diseases and have been implicated in the regulation of gene expression during adaptive evolution (Gatchel and Zoghbi 2005; Kashi and King 2006), which has led to increasing interest in their mutational dynamics. However, many aspects of microsatellite evolution are still unclear, particularly the implications of microsatellite mutation and diversification on genome dynamics.

1.1.2. Studying microsatellite mutation – the genomic age

The approaches used for studying microsatellite evolution are diverse and include vector based experimental studies and knock-out systems in model organisms, the analysis of multiple allele transmissions and allele frequencies in pedigrees and populations, and phylogenetic investigations (for a summary see Vargas Jentzsch *et al.* 2008). Naturally, each of these methods has advantages and limitations over others (e.g. see Amos *et al.*

2003), but a major problem in many cases has been the difficulty to draw broader conclusions about microsatellite evolutionary processes on a genomic scale. The advancing availability of genomic data throughout the last decade has made *in silico* approaches, particularly comparative studies, increasingly popular and enabled researchers at least in part to overcome the previous limitations. Taxa and species specific differences and similarities in microsatellite distribution have now been established (Toth *et al.* 2000; Katti *et al.* 2001) and have greatly facilitated the development of microsatellite mutation models (Kruglyak *et al.* 1998; Kruglyak *et al.* 2000; Dieringer and Schlotterer 2003; Sainudiin *et al.* 2004). More importantly, combined with new data and insights gained from areas such as functional and structural genomics, genomic analyses of microsatellites may hold the key to elucidating the role(s) of these simple sequences.

The diversity of approaches available to study microsatellite evolution along with the complexity of microsatellite mutations (i.e. frequently varying mutation rates between species, loci and even alleles (for a review see Ellegren 2004)), has created a large and diverse body of literature that is marked by 'consensus and controversy' (Chambers and MacAvoy 2000; Schlotterer 2000; Buschiazzo and Gemmell 2006). A predominant example of this controversy is the inconsistency in the definition of a microsatellite throughout the literature (Chambers and MacAvoy 2000; Merkel and Gemmell 2008). Another concern is that although there are a considerable number of protocols available for laboratory techniques related to the use and study of microsatellites, with many of these evaluated rigorously over 20 years of experimentation, there are no or only very limited protocols available for bioinformatic investigation of microsatellites. Despite their common use, most microsatellite detecting tools, particularly the in-house scripts used by individual investigators and laboratories, are poorly described. With the increase in available genome sequence data, we expect a further increase in microsatellite studies, thus resolving this gap in knowledge and similar problems is mandatory to maximize the insights gained from bioinformatics approaches.

1.1.3. Genomic distribution of microsatellites

The list of species in which microsatellites have been mapped, isolated and/ or analyzed is extensive, spanning all domains of life (Field and Wills 1998; Toth *et al.* 2000; Trivedi 2006). Although some taxa specific trends have been reported, more commonly genomic microsatellite signatures are unique and often differ even between closely related species (Ross *et al.* 2003; Karaoglu *et al.* 2005). Primarily, microsatellites are non-coding, mostly AT-rich DNA sequences (Toth *et al.* 2000; Katti *et al.* 2001). They also appear in coding regions, albeit to a much lower extent and almost exclusively as tri-and hexanucleotides, due to the deleterious effects of frameshift mutations caused by other repeat units (Metzgar *et al.* 2000; Toth *et al.* 2000). For example, Toth *et al.* (2000) estimated microsatellite coverage (bp per megabase DNA) in eukaryotes to be less than 0.23% in coding regions and 0.23 – 1.45% in non-coding regions). Regardless of the major trends, actual microsatellite frequencies vary somewhat depending on how the search is conducted and the quality of the analyzed sequence (i.e. later studies have shown microsatellite coverage in vertebrates to be up to 5% (Warren *et al.* 2008)). Across all eukaryotic genomes poly(A) and poly(T) have been identified as the most common motifs, followed by $(AT)_n$ and $(AC)_n$ in plant/fungi and animal genomes, respectively (Toth *et al.* 2000; Morgante *et al.* 2002; Karaoglu *et al.* 2005). Conversely, poly(C), poly(G), and $(GC)_n$ are rare in most genomes (Toth *et al.* 2000; Katti *et al.* 2001; Morgante *et al.* 2002).

On an intra-genomic level microsatellite distribution has been studied to a much lesser extent. Overall, microsatellite frequency appears similar amongst chromosomes, despite smaller chromosomes and sex chromosomes frequently exhibiting heightened densities of microsatellites (Bachtrog *et al.* 1999; Consortium 2002; Subramanian *et al.* 2003; Vargas-Jentzsch, unpublished data). Within chromosomes, it seems that microsatellites occur at relatively constant intervals, apart from around telomeres where microsatellite frequencies show a slightly elevated frequency (International Mouse Genome Consortium, 2002). Microsatellites are also often found associated with transposable elements, i.e. *Alu, LINEs* and *SINEs* in humans or *mini-me* in dipterans (Nadir *et al.* 1996;

Wilder and Hollocher 2001). This association has been linked to their genomic origin wherein they either emerge from the poly(A) tails of retrotransposons or are dispersed in a primordial form as so called "proto-microsatellites" (Nadir *et al.* 1996; Wilder and Hollocher 2001). In contrast, certain genomic regions in plants that are enriched with LTR transposons, are depleted of microsatellites (Morgante *et al.* 2002). It has been hypothesized that the discrepancy in plants arises because LTR retrotransposons neither bear poly(A)-rich regions nor potential "proto-microsatellites", and the accumulation of microsatellites has not caught up with the very recent expansion of these regions (Morgante *et al.* 2002).

1.1.4. Mutation mechanism(s)

Box: Strand slippage

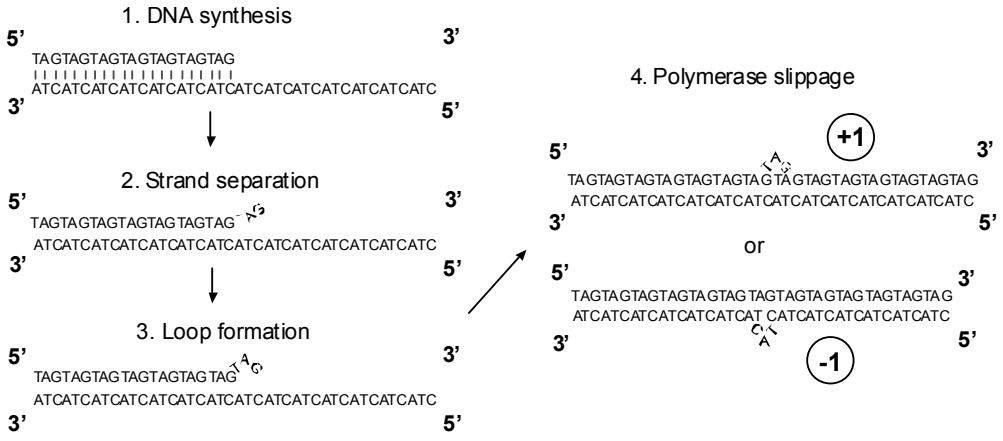


Figure shows strand slippage causing an increase (+1) or decrease (-1) in the number of repeats.

During strand synthesis the nascent strand dissociates from the template, forms a loop or hairpin and re-anneals out-of-phase with the template due to its repetitive nature. Subsequently, the DNA polymerase 'slips' past the loop and introduces an error, which, without further DNA repair activity, results in the addition or deletion of a repeat after reoccurring replication depending on whether the nascent or the template strand was involved in the loop formation, respectively.

Historically two mechanisms, strand slippage and recombination, have been proposed as the dominant agents of microsatellite mutation. Strand slippage (Levinson and Gutman 1987b) is widely accepted as the main cause for microsatellite instability and mutation, while recombination is thought to have only a minor contribution to these processes, if indeed it contributes at all (Sia *et al.* 1997; Schlotterer 2000; Ellegren 2004). Several observations support the claim: (i) no increase in microsatellite instability is observed in recombination deficient yeast and *E. coli* cells, and (ii) no difference is observed in the rates of microsatellite instability between mitotic and meiotic cells or autosomes and non-recombing sex chromosomes (e.g. human Y-chromosome) (Levinson and Gutman 1987b; Henderson and Petes 1992; Strand *et al.* 1993; Kayser *et al.* 2000).

Nevertheless, some authors have implied unconventional recombination events to explain cases of large deletions in disease causing trinucleotides and have proposed a model combining gene-conversion (no crossover) and strand slippage during strand synthesis (Richard and Paques 2000). Although the theory is mostly drawn from observation made with minisatellites (Jeffreys *et al.* 1994; Buard and Jeffreys 1997), a few studies in yeast (Richard and Dujon 2001) and bacteria (Jakupciak and Wells 1999), as well as in humans (Meservy *et al.* 2003), have detected an effect of recombination on microsatellites. In fact, weak correlations between the rate of recombination and microsatellite polymorphism have been shown by two independent studies in humans (Payseur and Nachman 2000; Bagschaw 2008). Alternatively, it has been suggested that microsatellite sequences themselves may stimulate recombination (Kirkpatrick *et al.* 1999; Gendrel *et al.* 2000).

Further, several studies have shown that hairpin structures may also promote trinucleotide instability in association with DNA repair and/ or replication fork stalling (for reviews see (Pearson *et al.* 2005; Wells *et al.* 2005; Mirkin 2007). Contrary to other observations (see below), mismatch repair (MMR) may facilitate microsatellite instability by binding to the hairpin structure, since some repeat expansions, requires a functional MMR system (Pearson *et al.* 2005). Repair and, or, recombination associated mutation mechanisms also pose an attractive explanation for the large numbers of multistep mutations observed

in species such as lizards (Gardner, *et al*. 2000), humans (Huang *et al.* 2002), *Drosophila* (Colson and Goldstein 1999), which are difficult to explain by replication slippage alone.

1.1.5. Microsatellite mutation models and life history

The peculiarities of microsatellite growth and decay have given rise to several mutation models over time which progressed as increasing data have become available. The original single step mutation model (SSM) was first introduced by Ohta and Kimura in 1973 and was resurrected in the early 1990's as it seemed to fit the observed mutational spectra for microsatellites better than the infinite alleles model (IAM) that had gained favour for dealing with allozyme data (Ohta and Kimura 1973; Valdes *et al.* 1993). The SMM predicts that alleles change by the addition or deletion of single repeats, with either event equally likely. Due to its simplicity it provided the basis for most estimators of genetic distances derived from microsatellite data used in population genetics; besides metrics derived from the IAM have proven more robust to the particularities of microsatellite dynamics (Goldstein and Schlötterer 1999; Balloux and Lugon-Moulin 2002). Later, the extended two-phase model (TPM), also incorporated the occurrence of multistep mutations, proposing distinctive probabilities for each type of mutation (Dirienzo *et al.* 1994). More recently, other models, though rarely applied in practice, have further invoked the rate and direction of length changes, motivated by observations of a length dependent mutation bias and the stabilizing influence of point mutations on microsatellite variability (e.g. Kruglyak *et al.* 1998; Xu *et al.* 2005, see below). In a comparison of existing models of microsatellite mutation that included most of the above described features, a proportional, linear-biased, one-phase model with a focal length towards which the mutational/ substitutional process is directed, emerged as best from all candidates (Sainudiin *et al.* 2004).

Combined with studies of microsatellite origin (and death), these theoretical approaches have resulted in a relatively comprehensive picture of microsatellite life history (Figure2).

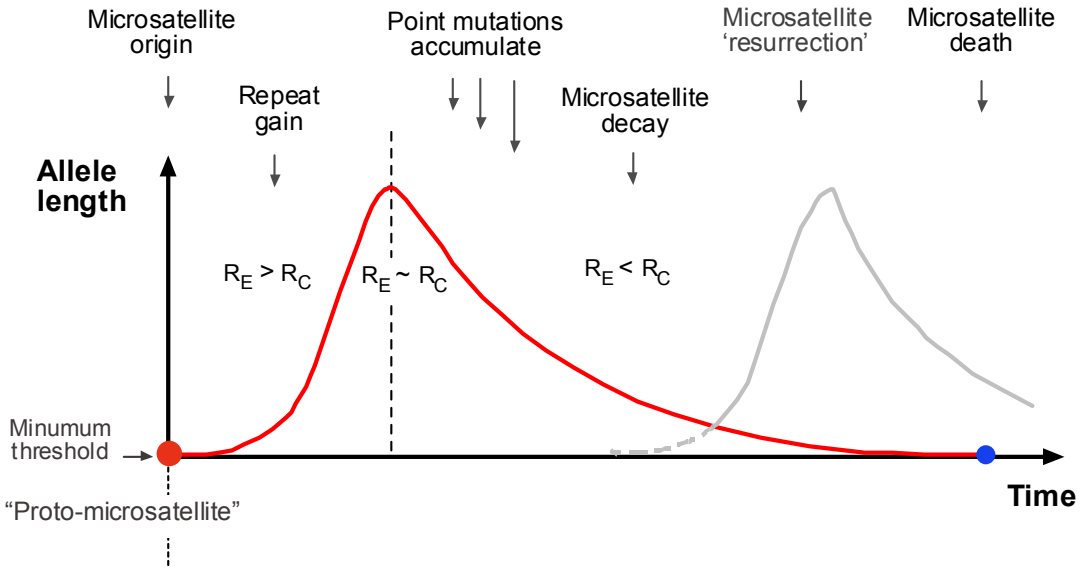


Figure 2. Microsatellite growth and decay.

R_E = rate of microsatellite expansions, R_C = rate of microsatellite contraction, $R_E \sim R_C$ = focal length

Although microsatellites may well emerge from the poly(A) tails of retrotransposons, or be dispersed in a primordial form ("protomicrosatellite"), more often they originate from random point mutations or small insertions/deletions (Nadir *et al.* 1996; Zhu *et al.* 2000; Wilder and Hollocher 2001). Once a certain length threshold is attained, strand slippage takes place, further expanding the repeat (Rose and Falush 1998). Initially the rate of array expansions is higher than the rate of array contractions, but eventually with increasing allele size the rate of contraction exceeds the rate of expansion leading to a decrease in array length (Xu *et al.* 2000). Meanwhile, point mutations slowly accumulate within the array, stabilizing the array at first but eventually degrading it unrecognizably into the genomic background (Kruglyak *et al.* 1998; Taylor *et al.* 1999; Rolfsmeier and Lahue 2000). Alternatively, a microsatellite may "resurrect" from its degraded state, and restart a new "cycle of life" at the phase of extensive expansions again (Chambers and MacAvoy 2000; Buschiazzo and Gemmell 2006).

The actual threshold size for slippage to occur, i.e. minimum array length or minimal repeat copy number, is subject to large controversy and causes frequent discrepancy amongst results (Merkel and Gemmell 2008). Studies in eukaryotes have inferred the minimum slippage length from the divergence between the expected length distribution

and the observed length distributions. The estimated point of divergence is ~8bp in yeast (Rose and Falush 1998), ~12bp for mono through trinucleotide, and 4 repeats or more for tetra through hexanucleotides repeats in humans (Lai and Sun 2003). The variance in these minimum slippage lengths indicates both locus- and species-specific differences in mutational processes that are potentially rooted in genome composition and/or evolutionary processes (Harr *et al.* 2002; Dieringer and Schlotterer 2003).

1.1.6. Factors influencing microsatellite instability

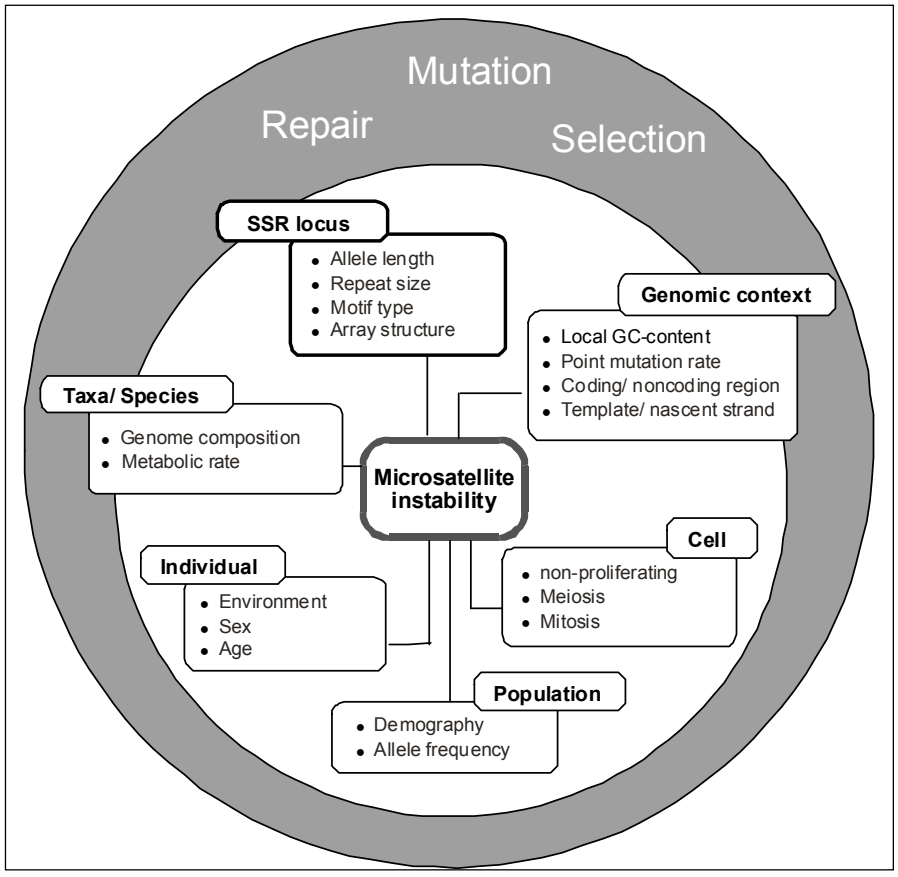


Figure 3. Overview of factors affecting microsatellite instability

Microsatellite mutation rate is affected by a multitude of factors that influence either the probability of generation of mutations or the efficiency of repair of these mutations

(Figure 3). These factors can be either direct properties of the microsatellite, such as allele length, repeat size, motif, and internal array structure, or extrinsic factors, that may act simultaneously on multiple levels (i.e. genomic context, cell, individual, population, taxa/species). However, on an evolutionary time scale, processes, such as selection and genetic drift, will further modify microsatellite variability and ultimately govern allele type and frequencies. In this section I review some of these influences.

Locus related factors

Frequent observations across multiple organisms in which long microsatellites are shown to be particularly polymorphic, have established a positive correlation between array length and mutation rate (Wierdl *et al.* 1997; Brinkmann *et al.* 1998; Primmer *et al.* 1998; Schug *et al.* 1998; Ellegren 2000). Supposedly, strand slippage occurs more often in longer arrays than shorter arrays, since the probability of undergoing strand misalignment increases with increasing locus repeat length (alternatively, some authors have argued that the key issue is the difference in allele size, provided gene conversion is a dominant factor in microsatellite mutation (Amos *et al.* 1996)). The direction of mutation that is the probability of expansion or contraction of the repeat array also seems to be dependent on length. Shorter repeat arrays show an expansion bias whereas longer arrays are biased towards contractions (Amos *et al.* 1996; Primmer *et al.* 1996; Wierdl *et al.* 1997; Ellegren 2000; Harr and Schlotterer 2000), which has been explained by the preferential repair of the newly synthesized strand (see below). Similarly, the highest slippage rates are reported for dinucleotides followed by tri-and tetranucleotides (Chakraborty *et al.* 1997; Schug *et al.* 1998); the explanation being that shorter motifs may allow more opportunities for loop formation, and have a lower dissociation energy than longer motifs. While these generalities seem to hold, different motif types vary in their ability to form different kinds of secondary structures and as such interfere with the dynamics of the mutational process and may lead to more frequent, larger multistep mutations that depart from the stepwise mutation model (for a review see Wells *et al.* 2005; Mirkin 2007). Finally, the internal architecture of the microsatellite, i.e. its sequence simplicity or complexity, also affects mutation. Mismatches within the array, reduce the strength of

strand mispairing and thus secondary structure formation, and have been frequently observed to lower microsatellite mutation rates (Goldstein and Clark 1995; Petes *et al.* 1997; Rolfsmeier *et al.* 2000). However, slippage can also remove introduced mismatches, countering their effects (Harr *et al.* 2000). Arrays composed of multiple motifs, that is compound microsatellites, exhibit much lower mutation rates than arrays composed of single motifs of the same length, although their individual segments mutate at rates higher than an isolated array of the same motif and length (Bull *et al.* 1999; Kayser *et al.* 2004).

DNA-Repair

Repair-deficient yeast mutants and colon cancers, that also have no functional mismatch repair (MMR), were the first indicators that MMR may have a role in microsatellite mutation since microsatellite instability was elevated by several orders of magnitude in those cells (Aaltonen *et al.* 1993; Strand *et al.* 1993; Wierdl *et al.* 1997). In comparison, yeast mutants deficient in exonucleolytic proof reading only showed a 5-10 fold increase in microsatellite instability (Kroutil *et al.* 1996). Mismatch repair maintains genome integrity by correcting single basepair mismatches and small loops in a substrate specific manner preferentially on the newly synthesised strand (Marra and Schar 1999). With respect to microsatellite instability, mismatch repair efficiency decreases with increasing unit size in yeast (Sia *et al.* 1997) and shows a bias amongst motifs (Harr *et al.* 2002). Mismatch repair also differs amongst different parts of the genome, between the leading and lagging strand, and has been suggested to cause species-specific differences in microsatellite distributions (Harr *et al.* 2002; Pavlov *et al.* 2003; Hawk *et al.* 2005).

Genomic context

In their role as genomic entities, microsatellites are likely to be affected by the properties of their genomic environment. Most noticeable, microsatellite type and distribution varies throughout the genome due to coding sequences (see above). However, flanking sequence GC-content, due to their strand separation properties, as well as local genomic

sequence composition and mutation rate have also been implied in the variability of microsatellites (Glenn *et al.* 1996; Brock *et al.* 1999; Santibanez-Koref *et al.* 2001; Dieringer and Schlotterer 2003; Pardi *et al.* 2005; Kelkar *et al.* 2008). The reports are nevertheless somewhat contradictory. On one hand, Brock et al (1999), studying human trinucleotide disease loci, and Pardi et al (2005) investigating characteristics of human $(AC)_n$ marker loci, both find a positive correlation between flanking sequence GC-content and microsatellite variability. On the other hand, Glenn et al (1996), utilizing dinucleotide loci in American alligators, detect a negative correlation between flanking sequence GC-content and allelic diversity, and Santibanez-Koref employing mouse-rat orthologous $(AC)_n$ loci detect no correlation between flanking sequence composition and array length at all. Furthermore, Dieringer and Schlotterer (2003) have shown an increase in microsatellite density for locally (1Mb) biased sequence composition (i.e. either elevated AT-content, or GC-content) across microsatellite distributions derived from 10 eukaryotic genomes. A recent study by Kelkar et al. (2008) showed negative correlations of microsatellite mutability with local GC-content (5Mb), particularly for dinucleotides, but no significant differences across different classes of isochores (data from human-chimp alignments). Such puzzling results could be simply explained by biases due to different sample sizes, but also may indicate, that factors influencing microsatellite variability act on different scales. In any case, more studies on those aspects are needed.

Species-specific characteristics are thought to manifest on a genomic level. For example, genomic sequence composition has been implicated in the overrepresentation of certain microsatellite motifs, e.g. the low GC-content in certain fungi genomes might explain the abundance of AT-rich motifs in those species (Lim *et al.* 2004). Further, the type and content of genomic features that either facilitate microsatellite emergence, such as transposable elements, or act against them, such as coding sequences, will, depending on their representation within a genome, further influence the frequencies of microsatellites.

Biological settings

Considering that DNA metabolic processes (replication, recombination, repair) modulate microsatellite mutation, it is conceivable that the biological settings of the cell (meiotic, mitotic, non-proliferating) and, or, organism (environment) are also relevant since they determine type and magnitude. For example, trinucleotide instability has been shown to be tissue specific (Martorell *et al.* 1997; Gomes-Pereira *et al.* 2001), and environmental factors such as radiation or oxidative stress, which induce DNA damage, have been reported to enhance microsatellite instability (Ellegren *et al.* 1997; Jackson *et al.* 1998; Hussein *et al.* 2005).

The sex and age of an individual also alter the probability of the transmission of a mutated allele into the next generation (Ellegren 2000). Males produce more germ cells through out their life time than females, i.e. they undergo more mitotic divisions, which is expected to cause a mutational bias (Brinkmann *et al.* 1998; Ellegren 2000) Sex differences will be pronounced in a species-specific manner, depending on the ratio between number of cells generated during spermatogenesis and oogenesis Finally, increased rate of DNA synthesis, i.e. cell divisions, in organisms with high metabolic rate could explain the high abundance of microsatellite observed in rodents compared to other mammals that exhibit slower metabolic rates (Martin and Palumbi 1993; Toth *et al.* 2000).

Selection, population demography and genetic drift

Under the premise that microsatellites are neutral markers, their variability will be affected by selection on a nearby region through processes like genetic hitchhiking or background selection. Under genetic hitchhiking neutral variants are fixed as a result of linkage to a beneficial mutation that is spreading through the population (Smith and Haigh 1974). This greatly reduces neutral variability in a genomic region, but also is expected to give rise to more alleles (i.e. recently derived alleles) than might be expected purely on the basis of gene diversity. Under background selection deleterious mutations and linked variants are continuously removed from the population (Charlesworth *et al.* 1993). For microsatellites these influences have been illustrated by several studies in

which microsatellite variation has been successfully used for hitchhiking mapping (for a review see (Schlotterer 2003)). Furthermore, whereas selection only affects certain regions of the genome, population demography (i.e. bottlenecks, population expansion or migration) acts on the entire genome. For example, bottlenecks, since they represent a reduction in effective population size, reduce the average levels of variability. Population expansion, on the other hand, may counterbalance the loss of alleles caused by random genetic drift.

1.1.7. Functional importance of microsatellites

Microsatellite variability has been proposed to be evolutionarily advantageous; variable loci could be selected for under the model of the "mutator phenotype" as a source for genetic diversity, and high frequencies of reversible mutations may act as genetic on/off-switches (Caporale 2003; Kashi and King 2006). A well known example is the reversible switching or phase variation of so called contingency genes found in many disease causing bacteria (for a recent review see (Moxon *et al.* 2006)). The expansion or deletion of a microsatellite located in coding regions disrupts the reading frame and stops protein synthesis. Reoccurring slippage can easily reverse the effect and regain the functional gene product. Such on- and off switching is commonly found in genes related to surface molecules which determine the pathogen's adherence to the host or, alternatively, it's susceptibility to the host's immune attack (Moxon *et al.* 1994).

Extensive studies of trinucleotide diseases further reveal that in addition to frameshift loss-of-function mutations, microsatellite expansions can be deleterious in untranslated regions (UTRs) and introns. For instance, most cases of fragile X syndrome are caused by the expansion of a CGG- repeat in the 5'-UTR of the fragile X mental retardation 1 gene (FMR1), that leads to transcriptional silencing and the loss of the gene product (FMRP) (Pearson *et al.* 2005). Further, Friedrich ataxia is rooted in the expansion of $(GAA)_n$ located in the first intron of the frataxin gene. Here, the mutation inhibits transcriptional elongation and reduces frataxin expression (Pearson *et al.* 2005). Finally, altered mRNA or protein function can lead to abnormal pathways and interactions and cause

pathogenesis in diseases such as Huntington disease, dystrophia myotonica and various forms of spinocerebellar ataxia (Gatchel and Zoghbi 2005; Pearson *et al.* 2005).

Variable microsatellites located within regulatory regions can also source profound phenotypic effects. Length variations may alter transcription via interruptions of the highly sensitive sterical interactions between transcription factors or by directly influencing the efficiency of transcription factor binding, sometimes even in a length depending manner (Kashi and King 2006). For example, the presence or absence of a compound microsatellite in the regulatory region of the vasopressin receptor (*avpr1a*) in voles has been linked to the social behaviour of individual species (Hammock and Young 2005). Similar repeat variability and effects are seen in the promoter regions of PAX-6B which is responsible for neurodevelopment and brain plasticity in humans (Okladnova *et al.* 1998) and the chicken malic enzyme which is the NADPH provider during lipogenesis (Xu and Goodridge 1998). Likewise, Fondon et al showed that the length ratio of two adjacent microsatellites in the runt-related transcription factor *Runx-2* correlated with the scull morphology in different dog breeds (Fondon and Garner 2004). Finally, Sawyer et al observed an influence of microsatellite length variants in the clock gene *period* of *Drosphila melanogaster*, which controls the fly's circadian cycle and temperature (Sawyer *et al.* 1997).

In principle, studies that relate certain microsatellite alleles to a phenotypic effect resemble linkage analyses for mapping quantitative trait loci (QTL) (Goldstein and Schlötterer 1999). However, since the mechanistic basis for these associations are not investigated (linkage is assumed), a more in depth analysis of QTLs could reveal that microsatellites are not only simple neutral linked variants, but the actual cause of the observed phenotypic variation.

1.2. Yeast as a model organism

Brewer's yeast (*Saccharomyces cerevisiae*) has accompanied human civilization for several thousand years having been utilized for wine making, brewing and baking, and more recently, has become one of the preeminent model organisms for eukaryotic

biology. *Saccharomyces cerevisiae*, strain S288C, has been a model organism for genetics and molecular biology for over half a century due to its ease of manipulation and genetic traceability. It is a single-cellular fungus with a short generation time (~90 minutes), inexpensive to grow and maintain, and stable both in diploid as well as in haploid state (Botstein *et al.* 1997). Its haploid genome is comparatively small (12.07Mbp) and packaged in to 16 chromosomes ranging between 0.2Mbp and 1.5Mbp in size (after Saccharomyces Genome Database, accessed 5th May 2008).

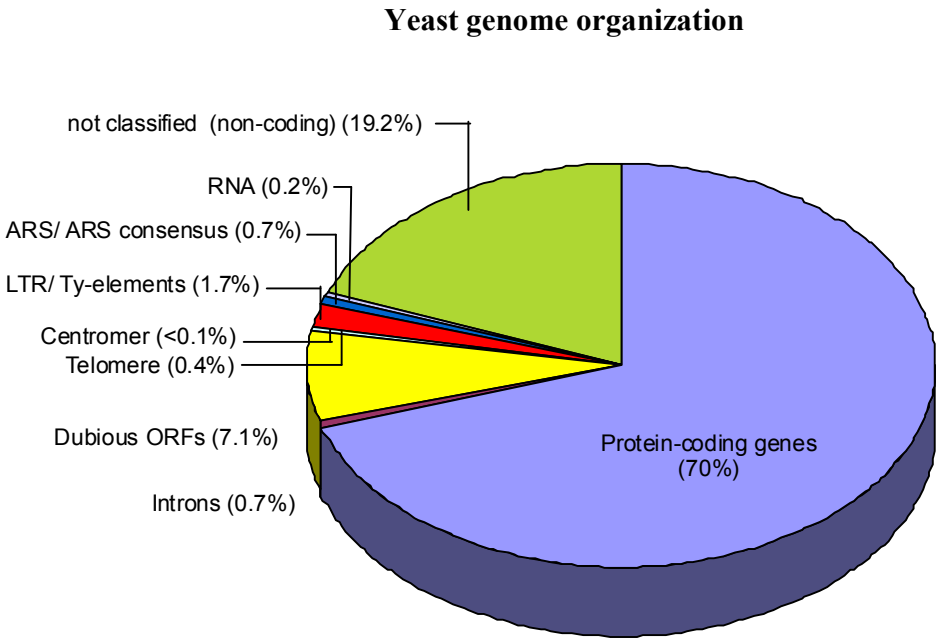


Figure 4: Yeast genome organization (annotations from SGD, as available 30/06/2008). Dubious ORFs = computationally predicted ORFs, but with no evident gene product; LTR = Long Terminal Repeats; Ty-elements = yeast specific transposons (Ty1-Ty5); ARS = autonomously replicating sequences (ARS consensus sequences function as essential components of replication origins in yeast). Introns are remarkably rare in the yeast genome, and so are LTR transposons when compared with, for example, their abundance in the human genome where they obtain 25.9% and 8.3% genome coverage, respectively (Lander *et al.* 2001).

Subsequently, yeast researchers pioneered a variety of functional genomics, through to the development of genome-scale experimental technologies (i.e. micro arrays and

derivatives), numerous analyses of global gene expression patterns and genome wide protein-protein interactions (for a review see Suter *et al.* 2006). This has yielded large amounts of genomic data (structural and functional), which can be freely accessed through numerous databases (e.g. SGD, *Yeast Protein Database* (YPD), *Comprehensive Yeast Genome Database* (CYGD) at the MIPS, and others). Its genomic features and wealth of annotations and resources make *S. cerevisiae* an excellent subject to study the specifics of coding and/or regulating microsatellites as well as the largely unknown contribution of intra-genomic influences to microsatellite evolution.

Advances in sequencing technology have further placed yeasts as phylum (= *Hemiascomycetes*) at the forefront of comparative genomics (Figure 5). Nearly two dozen yeast species have been partially or fully sequenced in recent years, greatly facilitating the study of the mechanisms of eukaryotic genome evolution, such as whole genome duplication and chromosomal rearrangements (Liti and Louis 2005; Dujon 2006).

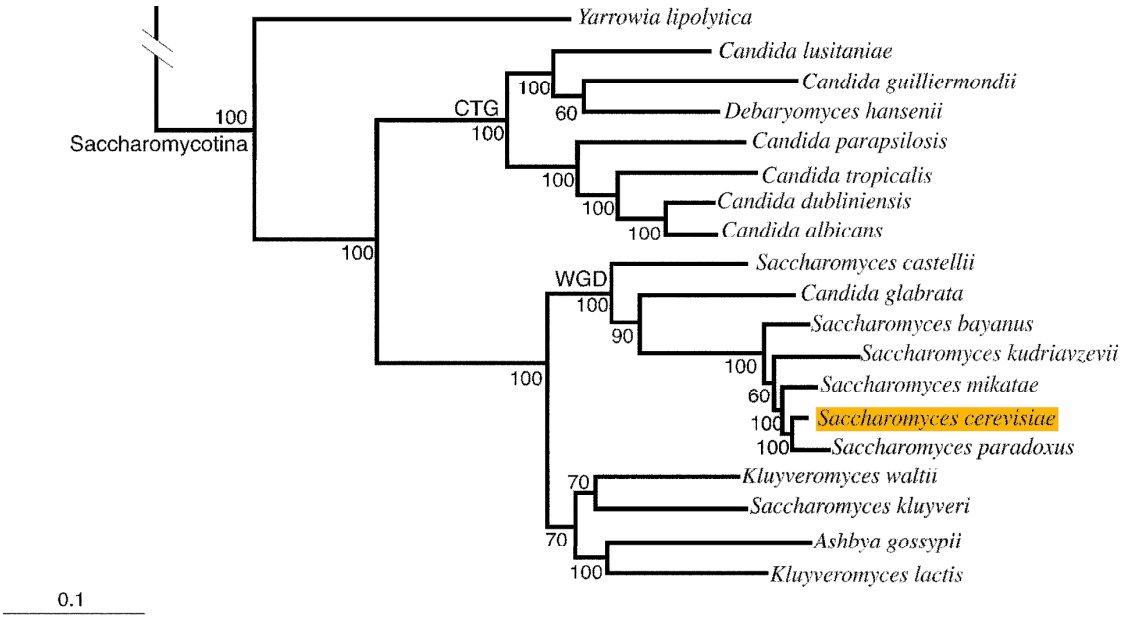


Figure 5. Maximum likelihood phylogeny of the Hemiascomycetous yeast species (sub-tree adopted from fungi super-tree after (Fitzpatrick *et al.* 2006)), WGD = whole genome duplication

The evolutionary range of the *Hemiascomycetes* has been estimated to resemble in size at least that of the chordates, enabling comparison of microsatellites across several hundred million years of evolution (the split from the filamentous fungi has been estimated ~300-400myr ago (Dujon 2006)). In addition, the recent genome resequencing of 40 *S. cerevisiae* (and 27 *S. paradoxus*) strains by the *Saccharomyces Genome Resequencing Project* (SGRP), provides extensive amounts of polymorphism data that allows microsatellite variability to be studied with respect to genomic variability.

At this point it hast be noted that first, microsatellites have been studied before in yeast (e.g. (Richard *et al.* 1999; Young *et al.* 2000)), and second, an increasing amount of genomic data is becoming available for other organisms and taxa, such as human, mammals, or *Drosophila* (e.g. see *Gene Expression Omnibus* at NCBI, http://www.ncbi.nlm.nih.gov/geo/). Nevertheless, we find that the combination and extend of the available data for yeast is exceptional and suits our study purposes more profoundly than any other organism.

1.3. Aims of this study

Microsatellite evolutionary dynamics are complex, and still remain undefined in many aspects. Factors that act on microsatellites within the genome (Ellegren 2004; Pearson *et al.* 2005; Kelkar *et al.* 2008) is one area of contention, and their potential functional roles, another, with an increasing number of studies showing support for the functionality of some microsatellites (Li *et al.* 2002; Li *et al.* 2004; Kashi and King 2006). The growing availability of functional and structural genomic data emerges as the key in resolving these issues.

In the first part of this dissertation, I address several problems that are associated with bioinformatic approaches used to detect microsatellites *in silico* from genomic data. A basic problem when characterizing microsatellite distributions is the exact definition of a microsatellite, particularly when making cross-species comparisons. Microsatellite definitions variously use the minimum array lengths required for a microsatellite, repeat

sizes, and the internal structure of the array, such as imperfect repetitions or multiple motif types (Chambers and MacAvoy 2000). In **Chapter 2**, I revisit the literature and trace the underlying reasons for this inconsistency. Utilizing a meta-analysis on microsatellite distributions in yeast, I further investigate the extent to which study results may diverge due different definitions and propose some guidelines that can be used to improve our ability to make comparisons among studies.

At the time of commencing this study, descriptions of the bioinformatic approaches used in genomic analyses to identify microsatellites in sequence data, were rare and mostly poor, despite their common application. It was difficult to find an appropriate tool without any prior knowledge of *in silico* methods. In **Chapter 3**, I describe the current range of microsatellite and/or short tandem repeat detecting tools, outline their structure, search strategy, efficiency/flexibility, and the specific utilities of each program to help users select the optimal tool for their purpose. Elaborating from Chapter 1, I also explore the possible detection biases for different tools (i.e. the preference for certain tools to detect certain types of microsatellites) and the influences of different parameter settings on the individual microsatellite searches.

Following up from the previous chapters, in the second part of my thesis, I use a bioinformatic approach, developed in light of my findings in Chapters 2 and 3, to study the intra-genomic influences on microsatellites using yeast as model organism. The wealth of structural and functional data available for yeast, allows simultaneous testing for a variety of factors, whilst excluding interferences from species-specific factors that have previously distorted conclusions drawn from similar studies across multiple species. In **Chapter 4**, I describe the distribution of microsatellites in the yeast genome and address several previously unresolved or uninvestigated hypotheses regarding the association of microsatellite with other genomic features, such as GC-content, meiotic double-strand breaks, SNP density and others. I employ two approaches, a generalized linear model (GLM) and wavelet transformation to further test whether the associations identified might act on different scales (≥1kb).

In **Chapter 5,** I further extend this approach and investigate the patterns of microsatellite conservation and polymorphism across 40 *S. cerevisiae* strains. Despite the general assumption that microsatellites are neutral markers, there is increasing evidence that microsatellites might be functionally important, particularly those located within genes or regulatory regions. The Gene Ontology of *S. cerevisiae* annotates a large number of genes, more than in almost all other species, most of which have been experimentally verified (Yon Rhee *et al.* 2008). It provides a unique opportunity to examine whether conserved and polymorphic microsatellite might be preferentially located in a group of genes associated with certain functions. First, I examine potential trends for locus conservation and polymorphism on a genomic level, i.e. whether microsatellites are preferentially conserved (or polymorphic) based on their chromosomal position, genomic fraction or inherent characteristics, such as repeat motif, repeat size or array length. Second, I investigate the functional potential of microsatellites located within and adjacent to genes and the selective constraint that may act upon these.

Chapter 2

Detecting Microsatellites in Genome Data: Variance in Definitions and Bioinformatic Approaches Cause Systematic Bias

2.1. Introduction

Microsatellites or short sequence/tandem repeats (SSRs/ STRs) are tandemly repeated DNA sequences of (commonly) 1-6bp length per repeat unit. Their high length polymorphism and abundance in all genomes make them the genetic marker of choice for a diverse range of applications spanning linkage analysis and genetic mapping through to forensics and ecological and evolutionary studies (Goldstein and Schlötterer 1999). Interest in microsatellite mutational dynamics is increasing, with significant interest emerging in the use of genomic data to investigate the evolution of these ubiquitous and useful sequences. To date, a significant number of studies have investigated microsatellite abundance in a range of species in order to examine the evolution of these simple sequences and infer their functional roles, if any, in gene regulation, genome structure etc. (Kashi and King 2006). Putative distribution biases have been investigated for introns, exons and intergenic regions as well as possible associations with other genomic elements, such as interspersed repeats (Arcot *et al.* 1995; Toth *et al.* 2000; Malpertuy *et al.* 2003; Li *et al.* 2004; Lim *et al.* 2004).

However, comparisons among large scale *in silico* genome studies, even from the same genomic data, are fraught with methodological bias. A recent paper by Leclerq *et al*. (2007) outlines significant differences among search algorithms based on intrinsic structure of the search algorithm and the parameter settings. We present a meta-analysis on microsatellite distribution in yeast as an example on how divergent study results can be in practice. We confirm Leclerq's (2007) findings, but more importantly we show that

the differences are rooted in a long-lived controversy, ever since microsatellites were first discovered 20 years ago; how exactly to define a microsatellite. Interspecies comparisons that derive from such different studies are particularly vulnerable to erroneous conclusions, and it is an intricate task to tease out the patterns of microsatellite evolution from those arising from study bias.

2.2. Methods

Table 1. Studies utilized in the meta-analysis. All studies report comparisons of microsatellite distribution pattern in yeast. Table shows (from left to right) study, algorithm or software employed, the type of repeat that was investigated (with respect to perfection/imperfection) and parameters that were implemented in the bioinformatics search, such as repeat size (mono-octanucleotide) and array length (minimum/ maximum threshold).

Study	Algorithm	Type of Repeat	Repeat Parameters
Field and Wills (1998)	PERL script - regular expression[1]	perfect repeats	all mononucleotides: 1–42bp repeat size: 2, 3, 4, 5, 6bp minimum length: 16, 24, 32, 40, 48, 56, 64bp
van Belkum *et al.* (1998)	C- script[2]	perfect repeats	repeat size: 1, 2, 3, 4, 5, 6, 7, 8bp minimum length: 10, 10, 18, 20, 18, 20, 21, 24bp
Katti Ranjekar and Gupta (2001)	C-script, - base-by-base search using adjacent sliding windows for alignments	imperfect repeats (mismatch every 10th nt)	repeat size: 1, 2, 3, 4bp minimum length: 20, 20, 21, 20bp
Dieringer and Schlötterer (2003)	C-script, - motif search for consecutive sequence stretches	perfect repeats (incl. partial copies)	repeat size: 1, 2, 3, 4bp minimum length: 2, 4, 6, 8bp maximum length: 20bp
Malpertuy, Dujon and Richard (2003)	TRF software (Benson 1999), -statistic/ heuristic approach	imperfect repeats (match: (+1) mismatch: (-2, -3, -4) indels: (-6, -9, -12))	pattern size: 2, 3, 4bp minimum length: 10, 15, 20bp maximum length: 20 repeats
Karaoglu, Lee and Meyer (2005)	PYTHON script	perfect repeats	pattern size: 1, 2, 3, 4, 5, 6bp minimum length: 10bp
Lim *et al.* (2004)	C ++ script, - base-by-base search using adjacent sliding windows for alignment	perfect repeats	pattern size: 1, 2, 3, 4, 5, 6bp minimum length: 5 repeats

[1] Personal communication, algorithm is now implemented as MsatFinder software (http://www.bioinf.ceh.ac.uk/ msatfinder/)

[2] The URL address given for the server was not valid anymore at the time of our study, no further information could be found

We undertook a meta-analysis of the published literature on microsatellite distribution in the yeast genome (*Saccharomyces cerevisiae*). The studies chosen are all comparisons of microsatellite distribution patterns (motif, size class, and array length) that include *S. cerevisiae* as one of the focal species, but differ in the approach and software used to detect microsatellite sequences (Table 1).

2.3. Results

All analyzed studies confirm unique species-specific motif distribution patterns and an over-representation of long arrays over short arrays, which is in concordance with current models of microsatellite evolution. However, we find large differences in the reported results (Figure 1). For example, Dieringer and Schlotterer (2003) report more repeats across all motif types than others, up to several magnitudes' difference. This study scored repeat frequencies (loci/Mbp) in the order of 10^4 for di- and trinucleotides and 10^3 for tetranucleotides, compared to 10^2 for dinucleotides and 10^1 for tri-and tetranucleotides, which are the next highest frequencies out of all other studies. Naturally, the number of repetitive sequences detected increases rapidly (exponentially) as the minimum array length decreases, which is particular apparent in the study by Dieringer and Schlotterer (2003) since the authors applied thresholds between 2 and 8bp.

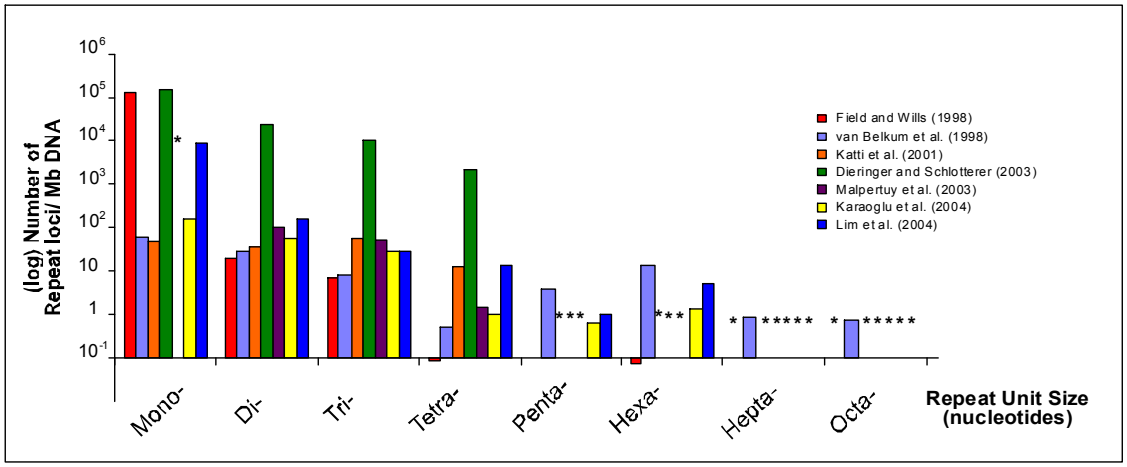


Figure 1. Microsatellite distribution in S. cerevisiae. Histogram shows the number of repeat loci per size class reported by each study. For details on parameter settings see Appendix Table 1); * = no data available

Subsequently, among all repeat sizes, mononucleotides are especially variable in the numbers of loci reported. We found frequency counts that ranged from a minimum of 46 loci/Mbp (Katti, Ranjekar, and Gupta 2001) to a maximum of 142,200 loci/Mbp (Dieringer and Schlotterer 2003). Further, the relative abundance of size classes also differs among studies. For example, all studies report mononucleotides as the most abundant size class with decreasing frequencies of longer repeat units, except Katti *et al.* (2001) who report the highest numbers for trinucleotides and van Belkum *et al.* (1998) who show an increased frequency for penta- and hexanucleotides.

2.4. Discussion

Given that the seven studies we examined have essentially analyzed the same genome data (small variations in build version not withstanding) for the same range of motifs, it is alarming to see such wide divergence in results. Here we discuss, that the crux of the problem derives from the different definitions of microsatellites used in each study. Differences in characteristics such as array length, unit size and purity inevitably transcribe into deviations in the parameter settings used in bioinformatics search tools, which subsequently lead to large discrepancies in results.

2.4.1. Microsatellite characteristics

Minimum array length

Historically, the preferred size for microsatellites selected as genetic markers has been a minimum of five repeats (Selkoe and Toonen 2006). However, the minimum array length required for strand slippage to occur is much lower. Rose and Falush (1998) determined a critical length at around eight nucleotides based on microsatellite distribution in yeast, while Lai and Sun (2003) approximated a minimum threshold of four copies for di-, tri-, tetra-, penta- and hexanucleotides and at least nine copies for mononucleotides for humans. In practice, however, the actual *in silico* detection of short repeats may be

restricted by the minimum resolution of the search algorithm, e.g. 10 or 11 nucleotides in the case of Tandem Repeats Finder (Benson 1999) used by Malpertuy *et al.* (2003). Repeat frequency increases exponentially with decreasing array length. Within our meta-analysis the differences in minimum cut-off length explain most of the variance: studies applying a low length threshold, e.g. in the case of mononucleotides around 2–5bp (Field and Wills 1998; Dieringer and Schlotterer 2003; Lim *et al.* 2004), harvest high repeat frequencies, whereas studies applying a higher threshold of 10 or 20bp report far fewer microsatellites (van Belkum *et al.* 1998; Katti *et al.* 2001; Karaoglu *et al.* 2005) (see Table 1).

Repeat unit size

Di-, tri- and tetranucleotide repeats dominate the literature because they have been found most frequently in the genome and are useful genetic markers (Jarne and Lagoda 1996). Mononucleotides, whilst common, have been largely avoided as they cause problems during amplification (Selkoe and Toonen 2006). However, from a mechanistic point of view, microsatellites are characterized by high levels of length polymorphism caused by DNA strand slippage, which can occur in repeat arrays composed of units that range from 1 to ~10bp in length (Levinson and Gutman 1987b; Jeffreys *et al.* 1994; Sia *et al.* 1997; Armour *et al.* 1999). Definitions of the motif length required to constitute a microsatellite vary in the literature: i.e. 1–6bp (Goldstein and Pollock 1997), 1–5bp (Chambers and MacAvoy 2000), 2–6bp (Schlotterer *et al.* 1998), or even 2–8bp (Armour *et al.* 1999). The same spread is reflected in our study survey: out of seven analyzed studies, one study excludes mononucleotide repeats (Malpertuy, Dujon, and Richard 2003), only four studies report numbers for penta- and hexanucleotides, and only one examines hepta- and octanucleotides (van Belkum et al. 1998) (see Table 1 for search parameters).

Purity and internal structure of the array

So far, the majority of *in silico* searches have investigated only perfect microsatellites as they are computationally easier to detect. However, perfect microsatellites are not the

only type of microsatellites. In fact, a repeat array might be classified as perfect (identical copies), imperfect (mismatches and indels are allowed) or compound/complex (array includes different motifs) (Chambers and MacAvoy 2000; Buschiazzo and Gemmell 2006). For most of the recent repeat detection tools, the level of imperfection can be varied as a parameter within the search. Despite this, (Katti *et al.* 2001) and (Malpertuy *et al.* 2003) are the only studies in our survey that allowed imperfections: a mismatch every 10th nucleotide, and succeeding mismatches after the first five perfect copies, respectively. While the available data do not allow us to detect a correlation between more or less stringent search criteria and high or low reported microsatellite frequencies, it appears logical that the inclusion or exclusion of imperfections in search parameters will influence the results of genomic comparisons.

2.4.2. Computational approach and genome build

There are additional, more subtle variables in the search that are rooted within the bioinformatic approach itself. Peculiarities of the underlying algorithm, such as combinatorial treatment of repeats in the identification procedure and/or redundancy filtering of overlaps or internal repetitions, may profoundly affect the overall pattern reported. Within our dataset, four studies (van Belkum *et al.* 1998; Katti *et al.* 2001; Malpertuy *et al.* 2003; Lim *et al.* 2004) apply the same minimum length threshold of 20bp in the case of tetranucleotides, but report frequencies of 0.5, 1.5, 12.6 and 13 repeats/ Mbp, respectively. Comparing the documentation for the search approaches (Table 1) suggests that studies using different algorithmic approaches report varying repeat frequencies. Unfortunately, details of parameter settings and the structure of the applied algorithm are not consistently published, thereby precluding detailed comparisons.

Different sequence builds and the inclusion of the mitochondrial genome (mtDNA) in the sequence analyzed can also contribute to variation in results. We ran TRF in default mode on three different *S. cerevisiae* genome builds and found no significant variation in the total numbers, types and distributions of the microsatellites reported (Appendix

Table1). However, a significantly higher frequency of microsatellites was detected within the mitochondrial genome compared to the nuclear genome (Appendix Figure 1) and the inclusion or exclusion of this genome in comparisons would result in a modest difference between studies.

2.5. Conclusion

The issue of how to exactly define a microsatellite is a long argued subject, upon that even today researchers have not reached consensus yet. Differences in parameters used in repeat detection, especially minimum array length, lead to large systematic biases in study results, where variations in microsatellite frequency can reach the extent of several magnitudes among studies even within the same genome.

Several authors have put forward microsatellite definitions, varying mainly based on the research background. First, describing types of repeats with respect to the degradation and complexity of the array subdivisions can be quite specific, such as in forensic and medicine (Urquhart *et al.* 1994), focusing on mutational behaviors of individual loci and alleles. We are predominately concerned with genomic analysis and propose therefore only three types of microsatellite including mono-hexanucleotides: perfect (repeat copies 100% identical), imperfect (mismatches and indels incorporated) and complex/ compound (consist of several motifs, potentially with mismatches). Second, minimum array length has been traditionally defined by the occurrence of strand slippage events and the extent of the resulting microsatellite polymorphism. This has led to analyses employing either stacked thresholds that depend on repeat size (for example see Table 1) or length classes, e.g. microsatellites class I: 12<20nt, microsatellite class II: >20nt (Temnykh *et al.* 2001).We suggest the following thresholds to start with, after Lai and Sun (2003): 12nt for mono-trinucleotides, 16nt for tetranucleotides, 20nt for pentanucleotids and 24nt for hexanucleotides. Absolute minimum thresholds for slippage events, tend to be group specific (between 8-15nt) and need to be adjusted individually

for each species to eliminate background noise, i.e. random occurrences of microsatellites, from true over- or under representation.

Ideally, future studies ensure that all data are gathered and analyzed in a consistent manner, which should enable a consensus approach to emerge within the literature. However, due to the potential intricacies of microsatellite distribution in different genomic architectures, this might not always be possible in an absolute manner. Therefore, we encourage all authors to report their parameter settings and algorithms in detail (including the underlying reasoning), to enable sensible comparisons across studies. The importance of the issue can not be emphasized enough in the genomic era, where cross-species comparisons are the tools of trade.

Chapter 3

Detecting Short Tandem Repeats from Genome Data: Opening the Software Black Box

3.1. Introduction

Microsatellites are short tandemly repeated DNA sequences (STR) of 1-6 bp unit length. Ubiquitously distributed in eukaryotic and prokaryotic genomes and highly polymorphic they rapidly became the current genetic marker of choice. Their usage is wide and includes genetic mapping, population genetic analysis, DNA forensics, and phylogenetics (Goldstein and Schlötterer 1999). More recently, microsatellite mutational dynamics have gained increasing interest as they have been shown to play a role in human genetic disorders (Pearson *et al.* 2005) and may have significant roles in the regulation of gene expression (Moxon and Wills 1999; Kashi and King 2006). For example, microsatellites have been found to be major effectors of morphological evolution in dogs and distinctive social behaviour in voles (Fondon and Garner 2004; Hammock and Young 2005).

With the sequencing of the first eukaryotic genome in 1996, the yeast *Saccharomyces cerevisiae* (Goffeau *et al.* 1996), a new *in silico* approach based on bioinformatic tools opened up for studying microsatellite evolution. Now, microsatellites could easily be detected from genomic data instead of using the cost- and labour-intensive laboratory approaches involving probe hybridization. To date, numerous algorithms and related software have been developed to explore microsatellite distribution in prokaryotes and eukaryotes, with investigations ranging from studies of regional distribution bias to putative association with genomic features (van Belkum *et al.* 1998; Katti *et al.* 2001; Li *et al.* 2002; Morgante *et al.* 2002; Dieringer and Schlotterer 2003; Li *et al.* 2004; Lim *et al.* 2004). These days, most sequence analysis packages or genome browsers incorporate

by default some form of tandem repeat finder, e.g. equicktandem and etandem at EMBOSS, repeat in the GCG-package and Tandem Repeat Finder (TRF) at UCSC (Benson 1999). Likewise so called repeat masking and low complexity filtering tools, such as RepeatMasker (Smit and Green 1996) or DUST/ SIMPLE (Hancock and Armstrong 1994; Alba *et al.* 2002), are now standard components of sequence similarity search tools, like BLAST and BLAST-like applications, to reduce redundancy and speed up genome-wide pattern match searches. Finally, several repeat specific databases have been established to serve as references for such diverging objectives as studying model organisms, e.g. TandemRepeatsDatabase TRDB (after Benson 1999) and EuMicrosatdb (Aishwarya *et al.* 2007), and DNA forensics, e.g. STRbase (Ruitberg *et al.* 2001). There are also numerous programs that detect repeats in protein sequences, some of which share feature with DNA-oriented detection algorithms (e.g. see Depledge and Dalby 2005; Kalita *et al.* 2006).

Two recent studies further denote the popularity of these tools. Leclercq *et al.* (2007) show a bias in repeat detection between algorithms, comparing some of the most commonly used tandem repeat finding programs, and Sharma *et al.* (2007) give a first overview over the available software for microsatellite detection while illustrating facets of microsatellite distribution in eukaryotic genomes. Nevertheless, for most biologists the variety of software tools is rather overwhelming and selecting an application appropriate for the question posed becomes a challenge. Here we describe the fundamental concepts implemented in short tandem repeat finding algorithms in order to provide a first practical guide to these commonly applied tools. We use examples from currently available software and discuss the utility of various applications for specific purposes. We see this information as an important step in moving biologists to develop selective approaches for microsatellite and repeat sequence detection, rather than the more common implementation of software as a mysterious black box.

3.2. Search algorithms

In simple terms, a repeat finder program consists of three components: a detection unit, a filter component, and the output compartment (Figure 1). The detection unit, harbouring the search algorithm is the core determinant of the overall time and space efficiency of the program. Based on certain selection criteria (statistics, scoring matrix) it detects patterns (motifs, repeats) specified under the users' input parameters. The resulting candidate repeats then undergo a filtering step to eliminate various types of redundancy. Outputs and utilities can vary widely between programs, i.e. including detailed information on the individual repeat, summary statistics, or even additional modules for subsequent analysis (primer design, clustering or alignment).

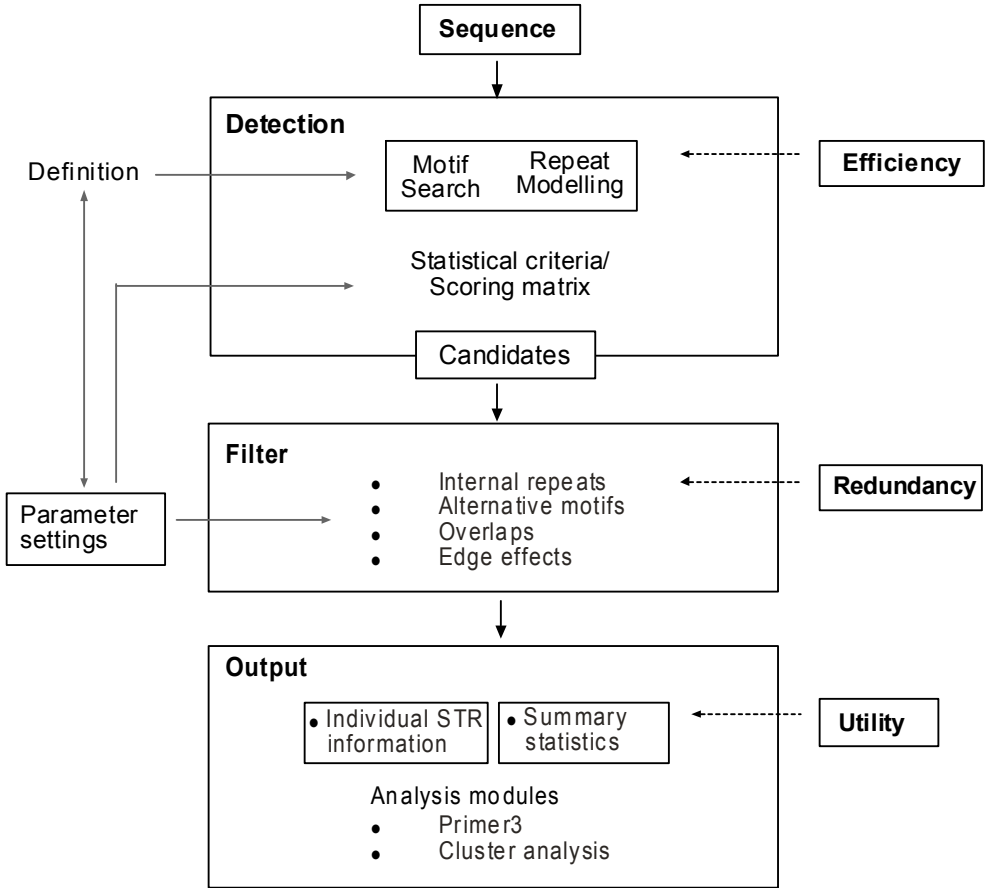


Figure 1. Schematic overview of a generic repeat finder algorithm.

3.2.1. Approaches

From a user's point of view, the identification of tandem repeats within a larger sequence takes two maxims: First, whether the search is going to be pattern specific, or unspecific (based on the repetitive nature of a sequence only); and second, what type of repeats will be searched for (perfect, imperfect or complex repeats) (see Table 1 for examples).

Table 1: Commonly employed terms for microsatellites/ short tandem repeats

Biological definition	**Mathematical/ Computational description**	**Features**	**Example**
Perfect	*Exact match*	100% identical copies	$(A)_n$, $(ATC)_n$
Imperfect	*Approximate - Hamming Distance (HD)*	substitutions (= mismatches)	$(AC)_nAT(AC)_m$
Interrupted*	*Approximate – Edit Distance (ED)*	substitutions, insertions, deletions (= interruptions)	$(ACG)_nT(ACG)_m$, $(AT)_nCGAG(AT)_m$
Compound/ Complex	*"Fuzzy"*	multiple motifs, periods, substitutions	$(ACG)_nT(TC)_m$, ATcgc|ATggc|ATtcc|ATcgg

* Interrupted repeats are often included in imperfect repeats.

From a programmer's point of view, the most straight forward approach to identify repeats is to search for specified sequences or motifs. In principle this can be achieved using any text editor, but practically, most searches investigate at least a set of motifs in very large texts, i.e. whole genomes. For some applications, like TROLL (Castelo *et al.* 2002), an application based on the Aho-Corasick algorithm (Aho and Corasick 1975), the user can provide a list of motifs in a separate input file which are then searched against the query all at once. Similarly, but based on a local alignment strategy, RepeatMasker (Smit and Green 1996) uses a list of pre-selected common motifs, stored in a reference database RepBase (Jurka *et al.* 2005), to scan a query for these sequences. Here the reference pattern is aligned along a genomic sequence implementing a scoring matrix. If a match is encountered, the adjacent sequences are aligned and subsequently masked if they exceed a certain threshold. Both programs are effective in detecting a defined set of

patterns in a sequence and are highly suitable for selective motif searches, but these are not effective substitutes for more comprehensive search tools (e.g. TRF (Benson 1999) or Sputnik (Abajian 1994), see below) that can be used for example to estimate genome wide repeat content.

Regular expressions are an efficient and hence popular way to search for repeats of a certain size and a large number of patterns. A regular expression describes a set of strings or patterns according to certain rules, such as the incorporation of wildcards into the motif at a fixed frequency. A variety of software languages accommodate regular expressions in their syntax, but due to its powerful inbuilt regular expression search engine *regex* and its text processing capabilities, many repeat detecting algorithms have been written in Perl, such as MsatFinder by Thurston and Field (2005), SSRIT (Temnykh *et al.* 2001) and MISA (Thiel *et al.* 2003). Msatfinder even employs regular expression searches at various levels of speed and accuracy: (i) fast regular expression (sequence is searched only once) and (ii) regular expression (sequence is searched several times); the first variant being a faster but less precise search and the second variant being slightly slower but more accurate in detection.

The first combinatorial approach to identify microsatellites/ short tandem repeats based only on repeat size, was implemented in the program Sputnik in 1994 (Abajian 1994). Sputnik employs a recursive algorithm using sliding windows to detect repeats of 1-5bp length by scanning through the sequence one base at a time, and checking subsequent bases for repeats. Matches of adjacent windows are evaluated by a scoring matrix. Initial repeats are extended and reported as long as they meet the minimum threshold. Poly (Bizzaro and Marx 2003), uses a similar base-by-base search, but differs from Sputnik (Abajian 1994) by searching for all window sizes at once instead of only searching for one pattern size at a time. The algorithm constructs accretive windows at each base of the input sequence, starting with the minimum pattern size. If there is no exact match to the preceding window, the window size is increased. Alternatively, if the maximum pattern size is reached and no match is detected, the starting position of the window shifts to the next base. However, both programs do not appear to differ remarkably in their execution

times. Since its initial release, Sputnik has been modified several times to improve either search capacity or output flexibility (Morgante *et al.* 2002; La Rota *et al.* 2005). The latest development from the Sputnik family tree is SciRoKo (Kofler *et al.* 2007), an extremely flexible tool, that incorporates fixed mismatch penalties as well as variable penalties (i.e. motif length*X).

Most of the approaches outlined above only search for very short tandem repeats (microsatellites) and/ or employ very simple substitution models, if a substitution model is employed at all. However, as a consequence of the recognition of tandem repeats as an essential component of all genomes analyzed so far and the general observation that imperfect/ complex repeats are more prevalent than perfect repeats, a large number of algorithms have been developed that model tandem repeats by employing the distance criteria (i.e. repeat size) as part of the search matrix itself. Such tools allow users to search for repeat sizes larger than microsatellites (e.g. minisatellites 10bp - ~100bp) and to search for specific types or patterns of repetition (see Table 1).

Amongst these, TandemRepeatFinder (TRF) (Benson 1999) is probably the most common and widely used tool for finding tandem repeats and has provided the basis for many other such tools (Wexler *et al.* 2005; Boeva *et al.* 2006). Initially, the algorithm uses sliding windows to search for matching nucleotides separated by a common distance. Like the Smith-Waterman-algorithm (Smith and Waterman 1981) it requires only partial matches between copies, called *k-tuple* matches (seeds). For each *k-tuple* match, the distance information and location are stored in an index. To select relevant candidates from the list a variety of statistical criteria are applied, which themselves are derived from several probability distributions (pattern length; matching probability p_m, indel probability p_i; and tuple size k). The result is not an exhaustive search but a sufficient one that in a heuristic manner enables reasonably fast processing of very large datasets, such as mammalian genomes.

ATR-hunter by Wexler *et al.* (2005) takes a similar heuristic/statistical approach. In addition to indexing the distance and location of potential repeat copies, it utilizes a

quality vector to describe the type of repetition. Applying scorings for matches and gaps of individual segments it is possible to find approximate repeats based on different similarity measures. Whereas TRF uses an alignment of each repeat copy to a consensus sequence as similarity measurement, ATR-hunter scores mutations between neighbouring copies or alternatively, the average similarity between all copies of the array, making it more flexible in detecting various types of repeats (e.g. Table 1).

Other applications have extended the concept of imperfect tandem repeats even further. TandemSwan (Boeva *et al.* 2006) detects so called 'fuzzy' repeats, i.e. repeats that can differ in number of mismatches per copy, period and number of copies. Based on an autocorrelation analysis, adjacent windows are compared to each other. Each letter comparison of a neighbouring window receives a score and repeats are eventually identified via a minimum function. The actual output candidates are selected via *p*-value thresholds based on the level of divergence between copies and motif similarity. Similarly, Mreps (Kolpakov *et al.* 2003) detects repeats composed of different motifs but is based on a seed extension technique instead. Here, initially exact repeats are detected which are then, dependent on a resolution parameter set by the user, maximal extended. All discovered hits undergo extensive redundancy treatment (see below) and are statistically verified based on a real distribution in a random DNA sequence.

3.2.2. Redundancy

Increasing the complexity and sensitivity of repeat detection is usually paralleled by increased redundancy in the discovered repeats, and thus the complexity of the analysis filter generally increases with the complexity of the search engine. Filtering is crucial for removing redundant output and particularly vital for accurate counts. However, the necessity for repeat filtering, and more importantly the type of filtering, should be determined based on the biological significance and research focus.

For instance, duplicated motifs such as $(ATAT)_2$ instead of $(AT)_4$, and permutations of the motif via alternative reading frames, such as AT versus TA, appear of no biological

difference and can easily be discarded as redundant. Whether AT or TA will be reported as a motif is subject to the neighbouring mismatches in the sequence and the threshold settings of the search tool. Generally, such location dependent redundancy filtering is achieved within the algorithm through a list or buffer where all repeat positions are recorded and from which eventually only a single hit per position is reported. Nevertheless, motif identification can be troublesome in the case of imperfect or very degenerate repeats. TRF (Benson 1999), for example, reports up to three possible motifs per locus allowing the user to manually check whether a motif has been correctly assigned to a repeat by the software, or not. This is potentially very useful when studying a particular motif type, but presents a major barrier to precise repeat counts and density estimates. Additional external redundancy filters may have to be applied if accurate counts are to be obtained (e.g. for genome wide microsatellite coverage). Alternatively, such as in Sputnik (Abajian 1994) and SciRoko (Kofler *et al.* 2007), permutations of a motif and the corresponding complementary motifs are grouped together in a natural sense (see Jin *et al.* 1994). The grouping of these motifs and their complementary motifs together has to be taken with caution if the research focus is on investigating microsatellite evolution, as some studies have shown strand preferences for certain motifs (Freudenreich *et al.* 1997; Morgante *et al.* 2002). Finally, merging of overlapping or adjacent repeats is yet another filter strategy, which is directed at certain repeat definitions, particularly compound or interrupted repeats, respectively. In some applications merging is optional (e.g. SciRoko (Abajian 1994), MsatFinder (Thurston and Field 2005)), but in others, such as Mreps (Kolpakov *et al.* 2003) merging is an integral component of the program and constitutes an additional purification step after a relaxed search. Here again, precise frequency estimates are traded-off for accurate motif distributions, and the choice of filter (or program) has to be made with respect to study purpose.

For example, if one was interested in the distribution pattern of $(AC)_n$ across various genomes, its frequency could be underestimated by merging or grouping. Programs like Star (Delgrange and Rivals 2004), TROLL (Castelo *et al.* 2002) and IMEx (Mudunuri and Nagarajaram 2007) (pattern search optional) can eliminate inferences from other

motifs through a motif specific search.. Alternatively, the same information could be retrieved via summary statistics (see below), provided merging and grouping options can be modified in the filter settings, e.g. Msatfinder (Thurston and Field 2005) or SciRoko (Kofler *et al.* 2007). On the other hand, if one was interested in overall microsatellite frequencies, such as occurrences per megabase (loci) or genome wide coverage (nt), the merging of overlapping repeats is crucial while sorting of motifs becomes irrelevant.

3.2.3. Study bias – algorithms and parameter settings

Naturally, different approaches are likely to diverge slightly in their outcomes, and tandem repeat detecting software is no exception. Nevertheless, we recently conducted a meta-analysis on published microsatellite distribution in yeast (Merkel and Gemmell 2008) that showed a divergence of up to three orders of magnitude in the frequency of microsatellite motifs reported among seven studies. We showed that the observed discrepancies are predominantly due to different parameter settings between studies which themselves emerge from different definitions applied for microsatellites (e.g. minimum array length/ repeat number, motif length, perfection/ degeneration of the array). We further found a bias depending on the algorithm employed (Figure 2) mainly in number of repeats detected, size classes identified and length distribution. Complimentary findings have been reported by Leclercq *et al.* (2007). Here, the authors tested five repeat finding programs, namely TRF (Benson 1999), Sputnik (Abajian 1994), Mreps (Kolpakov *et al.* 2003), STAR (Delgrange and Rivals 2004) and RepeatMasker (Smit and Green 1996), across several eukaryotic genomes and found major divergence in the repeats detected depending on the program, and more significantly the parameter settings selected. For example the study shows, that, at extreme values Sputnik (Abajian 1994) detects an 80-fold amount of perfect repeats detected by RepeatMasker on human chromosome X, and TRF (Benson 1999) shows an

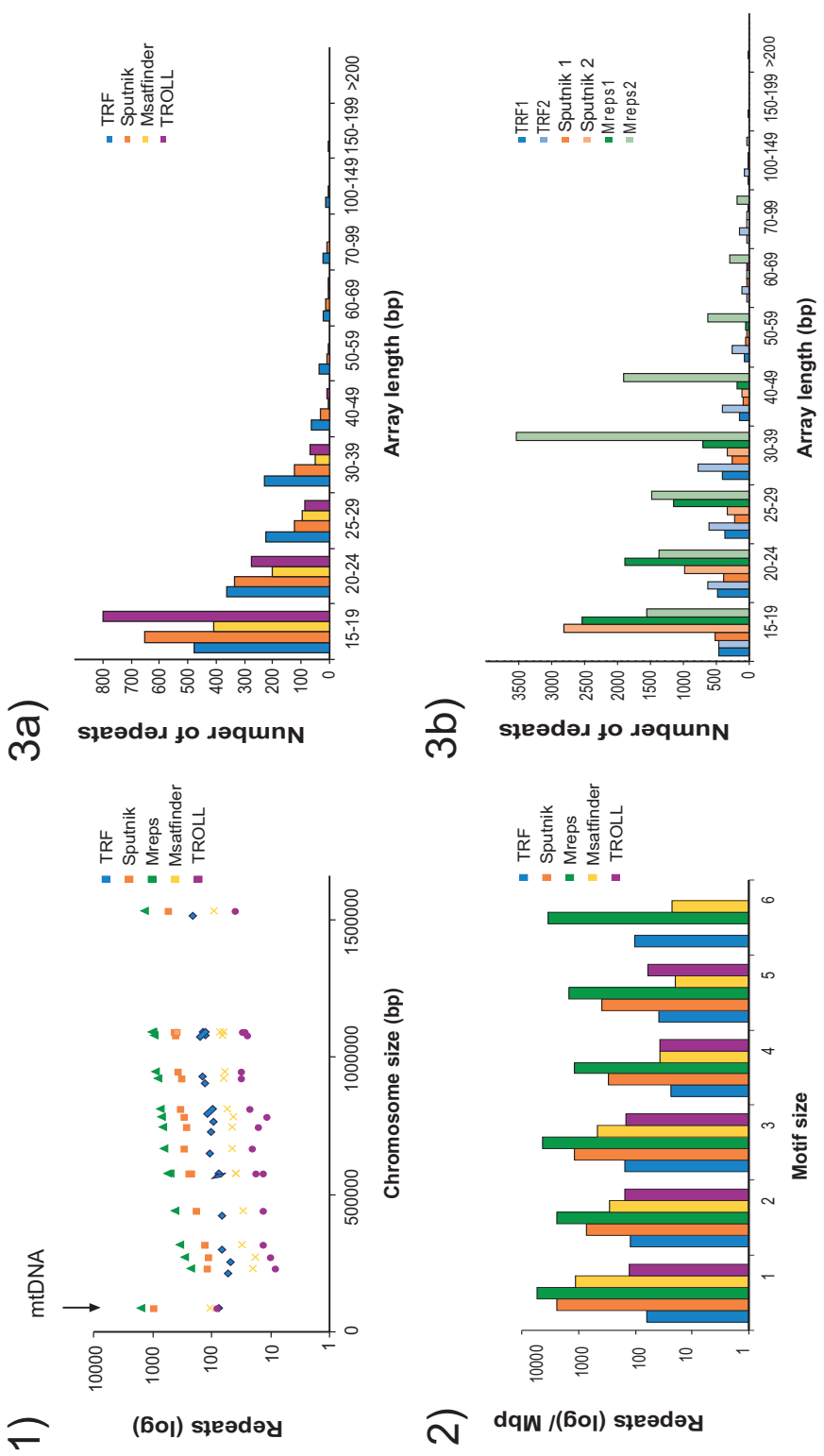


Figure 2: Comparison of microsatellite finding programs using the yeast genome. 1) Distribution of microsatellites across the mitochondrial genome and all chromosomes. Note the overabundance of repeats in mtDNA compared to nuclear DNA. 2) Distribution of repeat unit sizes. 3) Distribution of array length under a) stringent parameter settings which detects mostly perfect repeats) and b) relaxed parameter settings that allows for more imperfections in the repeat sequence). Noteworthy is the increased level of detection of intermediate length arrays for Mreps2, caused by Mreps inbuilt merging procedure. For 1) and 2) default parameters were used. For 3a) the parameter settings were: TRF: m =2, mism = 7, indels = 7, pm = 80, pi =10, score = 30; Sputnik: m = 1, mism = -6, score = 10; Mreps: res = 3, TROLL = default, MsatFinder = default. Parameter settings for 3b) are: TRF1: m =2, mism = 5, indels = 5, pm = 80, pi =10, score = 30; Sputnik1: m = 1, mism = -3, score = 12; Mreps1: res = 3; TRF2: m =2, mism = 3, indels = 5, pm = 80, pi =10, score = 30; Sputnik2: m = 1, mism = -3, score = 5; Mreps2: res = 1..Minimum array length (filtered) for all searches = 15nt, repeat size = 1-5bp.

61% increase in detections between two different alignment weights (2,7,7 and 2,3,5). Nevertheless, the observed biases were consistent across different genomes; hence, it seems there is no sequence specific program bias.

At a glance, such reports seem alarming and fundamentally question the accuracy of *in silico* microsatellite detection. Nevertheless, the underlying mechanics of the discrepancies can be traced. Considering algorithms implementing a scoring matrix for repetitive sequence identification, the standard parameters are minimum array length, minimum score and alignment weights. Minimum length is the most critical parameter for repeat detection, because short microsatellites are highly overrepresented in the genome. Hence, detections increase exponentially with decreasing minimum length. Threshold scores determine mean length and number of repeats detected but also influence the average degree of perfection within repeats, as imperfections lower the score (Leclercq *et al.* 2007). High threshold scores produce shorter and more perfect microsatellites, while lower threshold scores produce overall more, but on average longer and more imperfect repeats. In contrast, alignment weights (matches, mismatches, indels) predominantly extend or shorten already existing repeats, but only slightly increase the number of detections (Leclercq *et al.* 2007). Finally, threshold scores and alignment weights modulate the detected frequencies for different repeat size in quite a complex fashion, due to different size classes exhibiting unequal degrees of imperfection (Figure 3).

The individual search engine employed may also have an effect on the type of repeat detected with regards to average length and/ or the level of divergence in motif. TRF detects on average longer, but more imperfect repeats, whereas Sputnik detects shorter, but more perfect repeats (based on similar parameter settings and uniform divergence estimates) (Figure 2). This difference among the programs is likely due to TRF creating the repeat alignment based on a consensus sequence whereas Sputnik compares neighbouring copies to each other. Mreps (Kolpakov *et al.* 2003), which does not imply any minimum criteria for repeat identification such as score or length but a fixed seed size instead, shows no such bias, and detects repeats of equivalent degeneration

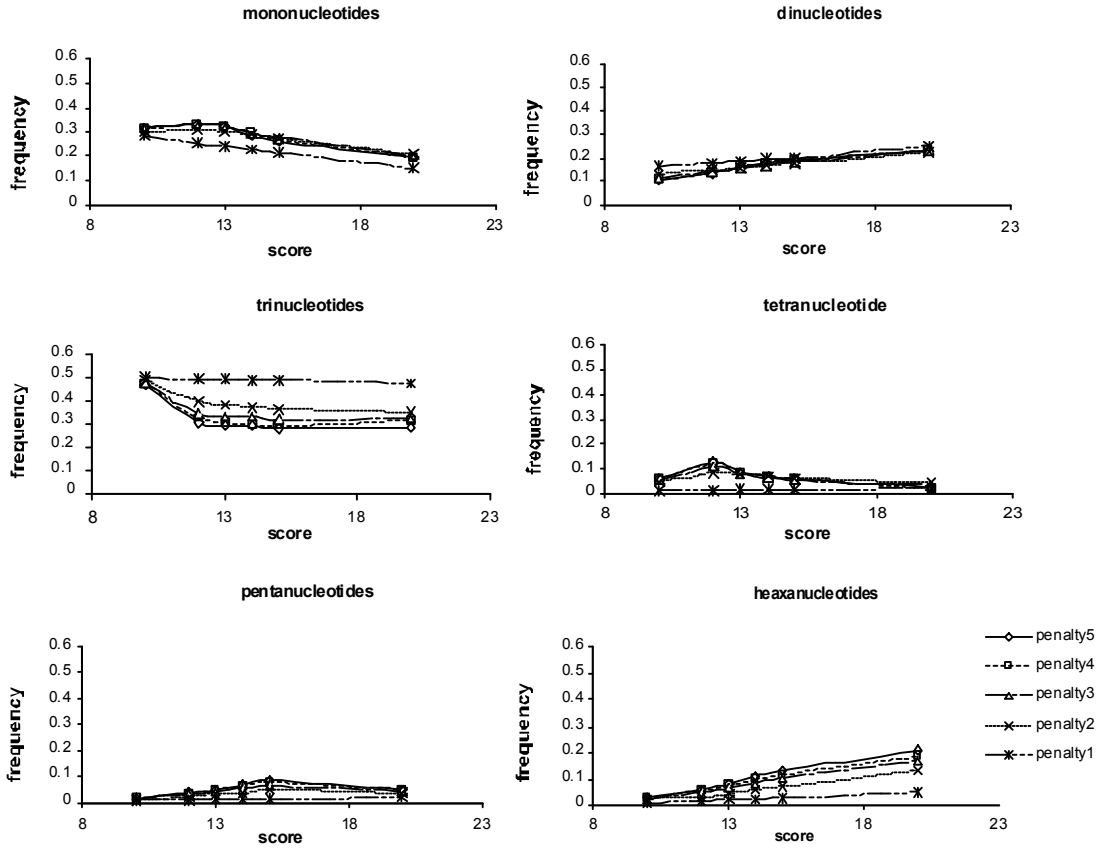


Figure 3: Influence of mismatch penalty and threshold score on different repeat sizes. High threshold scores and mismatch penalties increase di- and hexanucleotide frequency, as well as decreasing mononucleotide frequency. Tri- and tetranucleotide show temporary frequency changes at low scores, pentanucleotids at intermediate scores. Analysis was performed with SciRoKo (Kofler *et al.* 2007) on the nuclear *S.cerevisiae* genome only (score = minimum score, penalty = fixed mismatch penalty, minimum array length = 8 nt or 3 repeats)

regardless of their length (Figure 2). The longest and most divergent repeats are found by RepeatMasker (Smit and Green 1996) due to its pairwise alignment approach (Leclercq *et al.* 2007). Finally, repeat finders for only perfect repeats, like Msatfinder (Thurston and Field 2005) and TROLL (Castelo *et al.* 2002), identify naturally shorter repeats than other programs. Overall, the positions of most repeats overlap between programs in similar proportion as numbers of overall repeats detected increase. Still, some repeats are unique and combining search approaches can yield higher sensitivity.

Practically, problems are commonly encountered when searching for very short repeats. Those can only be detected at very low thresholds when using certain programs. But low thresholds create usually a lot of background noise, made up of highly degraded 'microsatellites' that appear to be closer to random sequences than microsatellites of biological significance. One way this can be avoided is using an additional program that has a higher fidelity for shorter repeats, and subsequently combing both results. However, as a backdrop, an additional filtering step becomes necessary to eliminate overlapping repeats. To avoid methodological biases and verify their results a few studies have employed multiple searches using a variety of parameter settings, different algorithms, or both (e.g. see (Naslund *et al.* 2005; O'Dushlaine *et al.* 2005; Kelkar *et al.* 2008)).

3.2.4. Efficiency

The issue of search efficiency becomes rapidly apparent when processing large datasets, such as whole genome data, on standard desktop machines or even laptops. The time and space requirements of a tandem repeat search algorithm are directly correlated with the intricacy of the search (Landau *et al.* 2001). So algorithms for detecting exact repeats have the shortest running times exhibiting a linear time progression followed by algorithms detecting approximate repeats under the Hamming distance model (logarithmic running times). The most computationally costly algorithms are those that detect approximate repeats under the edit distance model (quadratic running times). Many string matching algorithms use dynamic programming routinely as a technique to increase processing speed.

On a structural level, the number of computations can be efficiently reduced by pre-processing of either the input sequence or, in the case of a motif search, the pattern itself. For example, TROLL (Castelo *et al.* 2002) constructs in a pre-processing stage a keyword tree from the motif input file, which then can be used to search multiple sequences. A common technique for increased search speed is to transform the queried sequence into a complex data structure to enable fast look-ups. REPuter by Kurtz and

Schleiermacher (Kurtz *et al.* 2001) incorporates suffix trees to search not only for tandem repeats but also large interspersed repeats. Far more exotic, STAR by Delgrande and Rivals (Delgrange and Rivals 2004) utilizes methods from the field of data compression to simplify the queried sequence. The sequence, together with the recording of mutations/ alterations, is transformed into a significance distribution. Repeats are subsequently detected as maxima in the distribution. The authors claim that the method also has the advantage to allow pattern size independent scoring (see above).

3.2.5. Flexibility and utility

Parameter flexibility, output options and other utilities vary widely with the available software. As user knowledge, sophistication and needs increase, fixed or flexible parameters might be preferred. A number of programs offer besides the default settings a hierarchy of different search levels, such as basic, intermediate and advanced with increasing amounts of parameter flexibility, e.g. IMEx (Mudunuri and Nagarajaram 2007), ATRhunter (Wexler *et al.* 2005) or Msatfinder (Thurston and Field 2005).
With regards to the many fold output options and additional functions available, program selection at this point should be made with the prime focus on the downstream analysis requirements (Table 2). All programs report at a minimum genomic position and the type or sequence of the microsatellite. Most programs supply further information about the microsatellite such as length, size class, base count, flanking sequence, GC-content of flanking sequence, and, in the case of imperfect repeats, some measure of imperfection, i.e. matches, mismatches, indels, percentage perfection of or even an entropy indication of the sequence in TRF (Benson 1999). A few programs provide summary statistics, e.g. total count, base coverage/ density, average length, size class and motif abundance and some software also contain additional applications like Primer3 (Rozen and Skaletsky 2000), designing primers automatically from the flanking sequence or modules for cluster analysis (Table 3).

Table 2. Repeats, parameters and potential resources related to studies focusing on microsatellites and short tandem repeats

Study goal	**Type of repeats searched for**	**Parameter settings applied**	**Suggested resources***
Amplification of microsatellites (Primer design), Identifying polymorphic microsatellites (prediction or *in silico* allele scoring)	Polymorphic, i.e. long and perfect arrays	Stringent, high penalties, minimum score and thresholds	SSRprimer, IMex, MsatFinder, Misa, TRDB, VNTRfinder, PolyPredictR
Characterize genomic microsatellite distribution, Study microsatellite evolution	All types, specific motifs/ repeat unit sizes/ array lengths	Various (study specific)	SciRoko, Sputnik, IMex, TRF, Misa, MsatFinder
Estimating genomic microsatellite content	All types, non-redundant loci	Relaxed, low penalties, minimum score and thresholds	SciRoko, Sputnik, MsatFinder
Eliminate/ mask redundancy prone regions	All types	Highly relaxed	RepeatMasker, DUST, SIMPLE
Find STRs/VNTRs other than microsatellites, e.g. minisatellites	All types	Various (study specific)	TRF, Mreps, STRING, STAR, ATRhunter, TandemSwan, etandem, repeat

*N.B. this list of resources is not exhaustive

Hence, if the primary goal is primer design an application like IMex (Mudunuri and Nagarajaram 2007), MsatFinder (Thurston and Field 2005), SSRPrimer (Jewell *et al.* 2006) or Misa (Thiel *et al.* 2003), that includes a Primer3 module, is best suited to the task and depending on the amount of sequence data to be examined a stand-alone version might be chosen over the web-interface. Local stand-alone versions generally process large datasets much faster than web-based counter parts, whereas web-based versions spare the user the time-and resource consuming software install, and are sufficient for a small number of queries. On the other hand, if the research focuses on microsatellite distribution, such as for the purpose of characterizing microsatellite abundance or exploring genome architecture, the use of a standalone version providing a range of summary statistics detailed locus information is almost mandatory. SciRoko (Kofler *et al.* 2007), TRF (Benson 1999), Sputnik (Abajian 1994), and others (see Table 3) are all good

choices for such tasks. Some specialized applications, such as VNTRfinder and PolyPredictR, also allow the prediction of potential allele variations or directly evaluate these using either preset rules for polymorphism detection or a combination of TRF and sequence alignment methods (e-pcr or BLAST), respectively (Wren *et al.* 2000; Naslund *et al.* 2005; O'Dushlaine and Shields 2006). A last source of microsatellite data and analysis tools are the purpose built databases for repetitive sequences. Several large microsatellite databases have already been established by pre-screening whole genome sequences for repeats (see Table 3) and some genome browsers display microsatellite data routinely as an individual feature track, e.g. tracks in the UCSC genome browser created by RepeatMasker and TRF (http://genome.ucsc.edu/).

3.3. Conclusion

Applications for detecting microsatellites and other short tandem repeats are many and diverse. Key structural differences exist among these in terms of search engines, filter, and utilities. Program resolution varies, and a methodological bias is observed among programs that is especially pronounced when parameter settings vary. Caution has to be taken when choosing parameters if comparable results are to be obtained among studies. Microsatellite distribution in terms of frequency or coverage and over-/under-representation of certain characteristics, such as motifs, should be interpreted with respect to the approach, i.e. repeat type or definition, and candidate validation statistics/ filter. Finally, users may choose an application based on the repeat type, i.e. the repeat characteristic investigated, the efficiency and utility of the program, such as parameter flexibility, implementation (gui/ web) and modules available for additional analysis.

Table 3. Search tools used for STR detection, overview of features and properties

Program	Script	Operating System	URL	User Interface	Type(s) of Repeat	Summary Statistic/ Locus Info	Flanking Sequences	Pri-mer3	Reference
RepeatMasker	Perl	Unix	http://www.repeatmasker.org/	console/ web	perfect, imperfect (RepBase)	no	no	no	(Smit and Green 1996)
Msatfinder	Perl	Linux	http://www.bioinf.ceh.ac.uk/msatfinder/	console/ web	1-6bp perfect (imperfect, compound optional)	yes	yes	yes	(Thurston and Field 2005)
Misa	Perl	Linux	http://pgrc.ipk-gatersleben.de/misa/	console	1-6bp perfect (interrupted optional)	yes	yes	yes	(Thiel, Michalek et al. 2003)
SSRIT	Perl	Linux	http://www.gramene.org/db/searches/ssrtool	web	2-10bp perfect	no	no	no	(Temnykh, DeClerck et al. 2001)
Sputnik	C	Windows, Linux	http://espressosoftware.com/pages/sputnik.jsp	console	1-5bp perfect, imperfect (HD)	no	yes	no	(Abajian 1994)
Sputnik I	C	Windows, Linux	http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/	console	1-5bp perfect, imperfect (HD)	no	yes	no	(Morgante, Hanafey et al. 2002)
Sputnik II	C	Windows, Linux	http://cbi.labri.u-bordeaux.fr/outils/Pise/sputnik.html	console/ web	1-5bp perfect, imperfect (HD)	no	yes	no	(La Rota, Kantety et al. 2005)
Poly	Phyton	Linux	http://bioinformatics.org/poly/	console	1-4bp perfect	(yes)	no	no	(Bizzaro and Marx 2003)
TRF	?	Windows, Linux, Mac	http://tandem.bu.edu/trf/trf.html	gui/ web	1-2000bp imperfect (ED)	yes	no	no	(Benson 1999)

Program	Script	Operating System	URL	User Interface	Type(s) of Repeat	Summary Statistic/ Locus Info	Flanking Sequences	Pri-mer3	Reference
ATRhunter	C	Windows, Sun/Solaris, Linux	http://bioinfo.cs.technion.ac.il/atrhunter/ATRHunter.htm	web	1-500bp imperfect (ED)	no	no	no	(Wexler, Yakhini et al. 2005)
TandemSWAN	C, C++	Windows, Linux	http://strand.imb.ac.ru/swan/	console/ web	3-100bp fuzzy repeats	no	no	no	(Boeva, Regnier et al. 2006)
Mreps	ANSI C	Windows, Linux, Mac	http://bioinfo.lifl.fr/mreps/	console/ web	1-?bp fuzzy repeats	no	no	no	(Kolpakov, Bana et al. 2003)
STAR	?	Linux, Sun and Mac, Windows	http://atgc.lirmm.fr/star/	web	1-9bp imperfect (ED), motif specific	no	?	no	(Delgrange and Rivals 2004)
STRING	C	UNIX, Windows, Mac	http://www.caspur.it/~castri/STRING/index.htm	console	Imperfect (HD)	yes	?	no	(Parisi, De Fonzo et al. 2003)
TROLL	C++	Unix, Linux	http://finder.sourceforge.net	console	perfect (motif file)	no	no	no	(Castelo, Martins et al. 2002)
IMEx	C	Linux	http://203.197.254.154/IMEX/index.html	console/ web	1-6bp perfect, imperfect (ED)	yes	yes	yes	(Mudunuri and Nagarajaram 2007)
SciRoko	C#	Linux, Unix, Solaris, Free BSD, Mac	http://www.kofler.or.at/bioinformatics/SciRoKo/index.html	gui/ web	perfect, imperfect, compound	yes	yes	no	(Kofler, Schlotterer et al. 2007)
etandem/equicktandem	?C	Unix, Windows, Mac	http://emboss.sourceforge.net/	gui/ console	perfect, imperfect	no	no	no	EMBOSS
findpattern/repeat	?C	Unix, Windows, Mac	http://www.accelrys.com/products/gcg/index.html	gui	perfect, imperfect	no	no	no	GCG-package

Database	Algorithm	Genomic Data	URL	Features	Reference
Repbase	Reference database	Eukaryotes	http://www.girinst.org/repbase/update/index.html	Transposable elements, simple repeats, pseudogenes	(Jurka, Kapitonov et al. 2005)
TRDB	TRF	Eukaroytes (44)	http://tandem.bu.edu/cgi-bin/trdb/trdb.exe	Personal project space, flanking sequences, primer design, cluster analysis, TF binding site prediction, graphical representaion	(Gelfand 2006)
EuMicroSatdb	Misa	Eukaryotes (31)	http://veenuash.info/web/index.htm	Batch download, flanking sequence, compound microsatellites, genomic position (intron/exon, intergenic, upstream region)	(Aishwarya, Grover et al. 2007)
InSatdb	TRF	Insects (5)	http://210.212.212.8/PHP/INSATDB/home.php	Batch download, flanking sequence, compound microsatellites, GC-content, genomic position (intron/exon, intergenic, upstream region)	(Archak, Meduri et al. 2006)
ABCC GRID Database	STR finder	UCSC	http://grid.abcc.ncifcrf.gov.	Sequence, feature query, gff-format,view at UCSC	(Collins, Stephens et al. 2003)
Trbase	TRF	Human	http://trbase.ex.ac.uk/	Search for tandem repeats by gene/ desease association, genomic position	(Boby, Patch et al. 2005)
SSRPrimer & SSR Taxonomy Tree	Sputnik	GenBank	http://bioinformatics.pbcbasc.latrobe.edu.au/ssrdiscovery.html.	Primer design, taxonomy search, visualization	(Jewell, Robinson et al. 2006)
IMEx	IMEx	Prokaryotes, virus	http://203.197.254.154/IMEX/index.html	See IMEx	(Mudunuri and Nagarajaram 2007)
VNTRfinder & PolyPredictR	TRF	various	http://www.bioinformatics.ucsi.ie/vntrfinder/	Detection/scoring of homologous alleles is multiple sequences, polymorphism prediction, repeat type, flanking sequences, gene diversity/ heterozygosity	(O'Dushlaine and Shields 2006)

Chapter 4

Genomic Distribution of Microsatellites and their Association with Other Sequence Elements in Yeast

4.1. Introduction

Genome analyses and comparative genomics have taught us that genomes contain a lot less 'junk' than previously thought. Myriads of functional elements involved in gene expression have been identified in non-coding regions (Kellis *et al.* 2003; Mattick and Makunin 2006) along with the observation that DNA metabolic processes such as replication, recombination and nucleosome positioning are sequence dependent (Petes 2001; Nieduszynski *et al.* 2006; Segal *et al.* 2006). Repetitive sequences are a common feature of all genomes. Amongst these, microsatellites, a class of short (1-6bp) tandemly repeated motifs, can make up a substantial amount of the genome (e.g. up to 5% in mammals (Warren *et al.* 2008)). Microsatellites are ubiquitous in all known prokaryotic and eukaryotic genomes, which, together with an exceptionally high mutation rate, have made them genetic markers of choice throughout the last two decades for applications spanning genetic mapping through to population genetics, and DNA forensics (Goldstein and Schlötterer 1999). Whilst predominantly viewed as neutral markers an increasing number of studies provide evidence that these simple sequence repeats may sometimes diverge from that basic premise and may posses biological functions. The best studied of such examples are the excessive trinucleotide expansions involved in nearly 30 human heritable disorders (Gatchel and Zoghbi 2005; Mirkin 2007). Other instances demonstrate roles for microsatellites in the regulation of gene expression during adaptive evolution, where their frequent and reversible mutations have led some authors to designate them as "evolutionary tuning knobs" (Kashi and King 2006).

Intriguingly, microsatellites show a strong bias in their distribution within and amongst genomes (see *Chapter 1*), which has provoked numerous hypotheses regarding the association of these simple sequence repeats with other common intra-genomic elements. A well known case is the frequent association of microsatellites with retrotransposons in humans and dipterans, which has been largely linked with their origin within the genome (Nadir *et al.* 1996; Ramsay *et al.* 1999; Wilder and Hollocher 2001). (Under this scenario, microsatellites either emerge from a retrotransposon's 3' poly(A) tail after insertion (Nadir *et al.* 1996), or are dispersed with the retrotransposon in a primordial form as a 'proto-microsatellite' (two or three repeat copies (Wilder and Hollocher 2001))

Other less developed theories regarding putative roles of microsatellites spring from the heightened propensity of certain microsatellite sequences to form secondary structures or alter DNA structure as a result of their repetitive nature or their specific nucleotide composition (Baldi and Baisnee 2000; Mirkin 2007). These properties create an opportunity for microsatellites to interact with processes that strongly depend on DNA topology, such as replication, nucleosome positioning or transcription.

For example, replication origins (or autonomously replicating sequences in yeast) exhibit unique structural properties required for effective replication initiation, such as regions of low thermodynamic stability that facilitate strand separation. Ak and Benham (2005) showed recently that these regions are a result of regional DNA superhelical stresses. In theory, similarly to the connection found between negative DNA supercoiling and *in vitro* Z-DNA formation (Herbert and Rich 1996), such helical stresses could facilitate the stabilization of microsatellites forming alternative 3D-DNA structures in the nearby genomic neighbourhood which, in fact, has been shown for certain trinucleotides (Napierala *et al.* 2005). Conversely, other repetitive elements associated with DNA unwinding, such as $(ATTCT)_n$ involved in spinocerebellar ataxia type 10 (Liu, Bissler et al. 2007), might influence the structural properties of replication origin themselves. In both cases, microsatellite distribution is expected to differ significantly in these regions compared to their overall genomic distribution.

Further, transcription and transcriptional processes can be affected by microsatellites in several different ways. First, on an epigenetic level, effective transcription is strongly dependent on 'active' chromatin, i.e. accessibility to the DNA molecule for transcription factors. Accessibility in turn is determined by nucleosomes, the so called 'packaging units' of DNA, specifically their depletion (Brown 1999). Rigid structures like poly dA:dT tracts are well known to destabilize nucleosomes *in vivo* and *in vitro* (Iyer and Struhl 1995; Anderson and Widom 2001) and a few tri-, tetra- and pentanucleotide repeats have also been shown to deplete nucleosomes (Wang *et al.* 1996; Wang and Griffith 1996; Cao *et al.* 1998). Since poly(A/T) motifs represent the most commonly found microsatellites, we would expect a negative association between nucleosomes and microsatellites.

Second, several studies have reported that short tandem repeats can act as *cis*-regulatory elements in promoter regions (non-B DNA formation, protein binding sites). Amongst these studies are reports of microsatellites not only serving as binding sites for transcription factors (TF) (for example (Lafyatis *et al.* 1991; Chen and Roxby 1997)), but additionally having quantitative effects in TF binding efficiency based on repeat number (Kashi and King 2006). This modularity, based on the frequent repeat deletions or expansions, could provide a means to rapidly and efficiently adjust a quantitative trait (Kashi and King 2006). Despite these studies showing regional and gene specific support for functional roles for microsatellites, and although certain transcriptional regulators are known to bind repetitive motifs in yeast (Harbison *et al.* 2004), no such investigations have been made on a genomic scale for short tandem repeats.

Finally, some microsatellite sequences have been implicated in recombination, primarily as a consequence of their over representation at sites of high frequency recombination, so called hotspots, in yeast and human (Majewski and Ott 2000; Bagshaw *et al.* 2008). Experimental studies have demonstrated increased recombination frequencies after the insertion of a microsatellite, and additionally showed altered recombination frequencies depending on the number of repeat copies present (Kirkpatrick *et al.* 1999; Gendrel *et al.* 2000). However, a recent study by Buhler et al (Buhler *et al.* 2007) remapping double-strand breaks (DBS), the precursor of recombination events, found that these occur much

more often and much more evenly in the yeast genome than previously thought. Like previous studies the authors employed yeast mutants that had unrepaired DSBs. However, former studies where based on *rad50S* (strand exchange protein catalyzing the strand invasion step during repair) and related mutant phenotypes employing immunoprecipation to yield DNA stretches bound to Spo11 (homolog of the catalytic subunit of a type II DNA topoisomerase) (e.g. see Gerton et al, 2000). Buhler *et al* (2007) employed *dmc1* mutants (a different strand exchange protein) in order to map single strand DNA created at the break ends. Whereas in *rad50S* mutants DSBs are unrepaired and unprocessed with Spo11 remaining covalent attached to the strand ends, Spo11 is lacking from the further processed break sites in *dmc1Δ* mutants (for details see Buhler et al (2007)). Buhler et al (2007) observed five-times more DSBs created in *dmc1Δ* mutants than *rad50S* mutants that occur at much higher densities compared to previous reports (eg. a mean inter-hotspot distance of 9.5kb and 35kb, respectively), suggesting a higher involvement of Dmc1 in meiotic repair. The mechanistic details of the involvement of various proteins in DSB formation remain subject of further research. However, with respect to microsatellite affiliations, this new data implies that any proposed role for microsatellites in recombination may have to be re-examined as they may have wider effects than previously thought.

Here, for the first time, we analyse the genome wide implications of associations between microsatellites and genomic elements including: Ty-elements (LTR retrotransposons), tRNA, autonomously replicating sequences (ARS), ARS consensus sequences, meiotic double strand breaks (DSBs), regulatory sequences, centromeres, telomeres, introns and coding sequences (CDS). We chose the model organism *S. cerevisiae,* due to its high quality genome sequence and the wealth of genomic annotations associated with this sequence, most of which have been experimentally verified. In addition the recent release of 40 completely sequenced yeast strains by the *Saccharomyces Genome Resequencing Project* (SGRP, http://sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml) allowed us to further investigate the neutral behaviour of microsatellites in comparison to other regions of high polymorphism, such as high SNP density.

Although microsatellite distribution has been studied before in yeast (Field and Wills 1998; Katti *et al.* 2001; Dieringer and Schlotterer 2003; Malpertuy *et al.* 2003; Lim *et al.* 2004; Karaoglu *et al.* 2005), our study significantly improves on earlier analyses. First, the utilisation of different computational search approaches has led to rather inconsistent results among previous studies (Merkel and Gemmell 2008). Here we employ multiple searches to ensure consistency. Second, since those earlier studies were published new data has become available, which allows us to examine associations not tested previously. Third, new statistical approaches, such as wavelet analysis, provide a better analytical framework for identifying scale-specific effects among multiple genomic factors and the sequences of interest, potentially revealing associations and interactions at scales previously unconsidered.

4.2. Methods

Microsatellite search and analysis

We used the microsatellite detecting software SciRoko v3.3 (Kofler *et al.* 2007) to screen the entire nuclear genome of *Saccharomyces cerevisiae* for perfect (100% identical repeat copies) and imperfect repeats (including mismatches between copies) of at least 12nt in length. We conducted several independent searches under different parameter settings to exclude any study bias that is common in this kind of bioinformatic approach (see Appendix). For example, microsatellite counts can vary significantly with minimum thresholds and to a lesser extent with alignment weights and search algorithm (for a discussion on microsatellite definitions see (Merkel and Gemmell 2008). The AT-content of the repeat array and GC-content of the flanking sequence were calculated through in-house Perl scripts. The *S. cerevisiae* reference sequence was downloaded from SGD (ftp://genome-ftp.stanford.edu/pub/yeast/ as available November '05). Statistical analysis was carried out using R statistical software version 2.5.0. (available at http://cran.r-project.org/).

Genome wide microsatellite frequency and genomic features

We created an R-script to calculate microsatellite frequency in 10kb sliding windows (100bp overlap) along individual chromosomes (see also *Chapter 5*). Regions of high microsatellite frequency were subsequently extracted as intervals containing a number of microsatellites >2 standard deviations higher than the chromosomal average (mean microsatellite frequencies per window differ between individual chromosomes) and analysed for enrichment with known genomic features using a standard Chi-square test. Here, we chose sliding windows over intervals, as sliding windows allow greater accuracy in the identification of peaks in the distribution.

Next, correlations of microsatellite frequency with other known genomic elements were tested in 1kb and 10kb non-overlapping windows using a Generalized Linear Model. (The model was based on a quasi-Poisson distribution due to the non-normality of the data and over-dispersion (Quinn and Keough 2002). All features were counted as elements per window. Elements overlapping between windows were referred to the interval that had most coverage of the element (i.e. >50% of the element was located within the interval).

Genomic features were downloaded from *Saccharomyces* Genome Database (SGD) (ftp://genome-ftp.stanford.edu/pub/yeast/, November 2005). Promoter regions were not annotated as such, so we defined promoter regions as regions within a distance of 300bp from ORF start, as most regulatory elements are concentrated within that region (Harbison *et al.* 2004). Locations of regulatory elements are taken from SGD after Harbison *et al.* (2004). Nucleosome positions were kindly made available by Dr. E. Segal (Segal *et al.* 2006). The locations for double strand breaks (DSBs) initiating recombination were retrieved from Buhler *et al.* (2007). For DSBs, we used two data sets based on the mapping of meiotic ssDNA in (i) *rad50S* mutants with a hybridization signal >5, and (ii) in *dmc1*Δ mutants also with signal >5 (Buhler *et al.* 2007). GC-content was estimated for 1kb and 10kb intervals of DNA using an online-tool (http://tim.saraogtim.com/molbio/gccontent.php, March 2008). SNP density was derived

from genomic alignments of 40 sequenced *S. cerevisiae* strains to the reference strain (including gaps) by employing Perl scripts (data and script from SGRP, December 2007, http://sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml) (see also *Chapter 5*). Statistical analysis was carried out using R statistical software v2.5.0. (Ihaka and Gentleman 1996) (http://cran.r-project.org/) and coverage of genomic intervals was calculated using the Galaxy web-server (Giardine *et al.* 2005) (http://g2.trac.bx.psu.edu/).

Wavelet analysis

Wavelet analysis was undertaken using R-scripts derived from (Spencer *et al.* 2006) who used this approach to study the influence of recombination on human diversity. Since wavelet decomposition is carried out in fragments of 2kb size, we analyzed each chromosome twice to cover the entire chromosome and create replicates. For each chromosome, we analyzed the maximum available fragment size possible, once initiated from the chromosome start and once initiated from the chromosome end (for example: Chr1: 1-128000bp; and Chr1: 128001-230000bp). Microsatellites and feature abundance were measured in nucleotide coverage. All data was inspected for normality prior to analysis and log-transformed as required.

4.3. Results

4.3.1. Locus related patterns

In total, we detected 2732 perfect repeats with an average length of 16.27nt covering 0.35% of the genome and, depending on parameter settings, between 1922–2642 imperfect repeats with average lengths around 24.78–33.65nt (Table 1). The majority (>50%) of microsatellites in the yeast genome are AT-rich (>90% AT-content), are found predominantly in non-coding regions, and exhibit a lower GC-content in their flanking sequence than their coding counterpart (perfect: 1729 loci; imperfect: 1227–1856 loci) (see Appendix).

Table 1. SSR detected under various parameter settings using SciRoko (Kofler *et al.* 2007)

Parameter setting	SSRs (counts/Mbp)	Genomic coverage ‰ (coding, non-coding)	Average length (Stdev.)
Perfect repeats: Score: 12, min. length: 12 (1), 12 (2),12(3), 16(4), 20(5), 24(6)	2732 (226.33)	3.46 (1.88, 8.3)	16.27 (6.28)
Imperfect repeats (14n4): Score: **14**, fixed pe**n**alty: **4**, min. seed length: 8	2098 (173.81)	4.13 (2.4, 9.13)	24.78 (17.78)
Imperfect repeats: Score: **15**, fixed pe**n**alty:**2**, min. seed length:8	2642 (218.88)	7.15 (4.51, 14.72)	33.65 (21.93)
Imperfect repeats: Score: **15**, **v**ariable pe**n**alty: **1**, min seed length: 10	2507 (207.69)	6.84 (3.06, 17.34)	30.72 (17.42)
Imperfect repeats: Score: **15**, **v**ariable pe**n**alty: **2**, min seed length: 9	1922 (159.23)	3.74 (1.77, 9.3)	25.80 (16.46)

Genomic fraction, i.e. coding versus non-coding regions, is a strong determinant for locus related factors (Figure 1). In coding regions the AT-content of the repeat array is positively correlated with the AT-content of its flanking sequence, but not within non-coding regions. In other words, the array composition appears more similar to the composition of the flanking sequence in coding regions than in non-coding regions.

We further observe a significant negative correlation between array length and repeat AT-content, as well as between mismatches and repeat AT-content. This negative correlation exists in non-coding regions, but not in coding regions. This again confirms that microsatellites in non-coding regions are AT-enriched, but also shows that the introduced mismatches are predominantly G/C indels/substitutions. Array length and microsatellites show similar correlations. However, comparing the correlations that we observe within the individual fraction with those from the genomic distribution, we find that microsatellite distribution on a genome wide scale is dominated by the characteristics of microsatellites in non-coding regions.

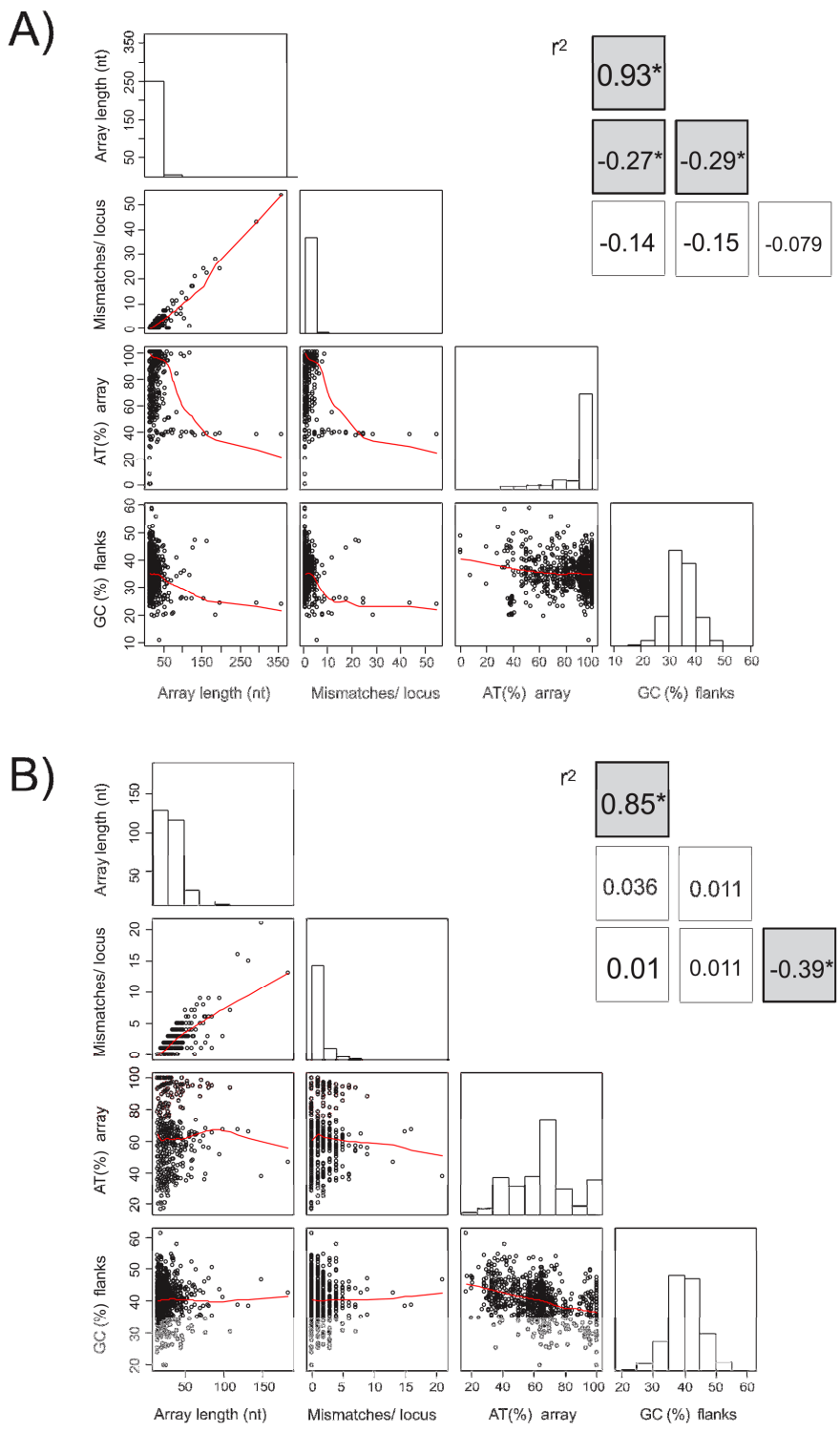


Figure 1: Imperfect (14n4) microsatellites in (A) non-coding and (B) coding regions of yeast. r^2 value shown,* indicates significant correlations with $p<0.001$ (Kendall's rank test for pairwise correlations).

Amongst repeat size classes, mononucleotide and trinucleotide repeats are by far the most abundant repeat size classes followed by (in descending order) di-, hexa-, penta- and tetranucleotides (Figure 2). Frequencies differ dramatically between genomic fractions. In non-coding regions microsatellite abundance diminishes exponentially with increasing motif length, whereas microsatellites in coding regions are almost exclusively trinucleotides.

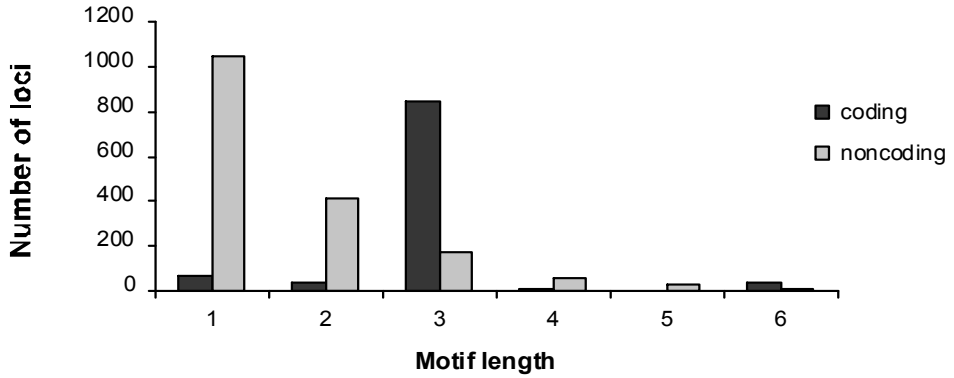


Figure 2. Distribution of repeat size classes for perfect microsatellites in different genomic fractions

Concordantly with the GC-bias of coding sequences, GC-content for microsatellites decreases with increasing unit size. Perfect mononucleotides are almost exclusively poly(A) and poly(T) motifs and trinucleotides show an exceptional GC-enrichment. Patterns for both perfect and imperfect repeats are very similar, although under relaxed (= imperfect) parameter settings the abundance of trinucleotides can exceed mononucleotides, and overall mono-and dinucleotides become more GC-enriched. Also the absolute number of detections strongly depends on the minimum length threshold set (see Appendix Table 1, Figure 1).

After poly(A/T), $(AT)_n$ and $(AAG/CTT)_n$ are the most common motifs within the yeast genome. Poly$(G/C)_n$ and $(GCC/GGC)_n$ are very rare, and $(GC)_n$ is not present at all. In total, the longest array in both coding and non coding regions is $(AAT/TTA)_n$ with a

maximum length of 65bp and 108bp, respectively. However, $(ACACCC)_n$ are the longest arrays on average. The motif distribution differs remarkably between non-coding and coding regions (Table 2).

Table 2. Microsatellite motifs for perfect repeats in non-coding and coding regions. Shown are number of loci, maximum length (in italics), and mean length per motif. All motif permutations and their complements are grouped together.

Motif	Non-coding regions			Coding regions		
	Count	Max. length	Mean length	Count	Max. length	Mean length
A	1044	42	14.9	72	32	14.5
AT	353	41	18.2	27	29	16.9
AAT	101	108	17.3	109	65	17.4
AC	45	62	17.3	9	41	23.4
AAG	27	20	13.6	223	48	15.2
AAAT	24	54	20.3	2	22	19.0
AAC	18	39	17.1	155	63	16.5
AG	13	64	19.7	2	21	18.0
ATAC	11	27	18.8	2	26	21.0
ACT	9	30	17.1	18	22	13.7
ATC	9	36	16.7	133	64	15.6
AGC	7	17	14.1	101	35	16.2
AAAG	6	23	19.0	1	17	17.0
ACACCC	6	43	31.3	1	30	30.0
ATCTC	6	35	24.5	0		
AAAAG	5	25	22.0	0		
AAAC	4	23	20.3	0		
C	4	19	15.8	0		
AAAAAC	3	49	34.0	0		
ACG	3	17	13.7	30	41	16.6
ATAG	3	19	18.7			
AAAAC	2	27	24.0	2	28	27.0
AAATT	2	26	24.0	0		
AATC	2	18	18.0	0		
AATG	2	18	17.0	1	17	17.0
ACACC	2	22	21.5	0		
AAAAAG	1	28	28.0	2	24	24.0
AAAAT	1	20	20.0	0		
AAACG	1	22	22.0	0		
AAATG	1	20	20.0	0		
AACAAG	1	32	32.0	1	26	26.0
AACAC	1	31	31.0	0		
AACAGC	1	24	24.0	10	33	28.7
AACAT	1	22	22.0	0		
AAGAT	1	22	22.0	0		
AAGG	1	17	17.0	0		

AATAC	1	25	25.0	0		
AATAT	1	20	20.0	1	20	20.0
AATCAT	1	37	37.0	1	30	30.0
AATT	1	17	17.0	0		
ACC	1	13	13.0	39	24	13.5
ACGT	1	18	18.0	0		
AGG	1	12	12.0	35	29	14.6
ATGC	1	16	16.0	0		
AAAGTG	0			1	24	24.0
AAATGC	0			1	27	27.0
AACATG	0			1	37	37.0
AACGAC	0			1	25	25.0
AACTCG	0			1	24	24.0
AAGACG	0			1	29	29.0
AAGAGG	0			1	25	25.0
AAGATG	0			3	39	30.0
AAGCAC	0			2	32	29.5
AAGCTC	0			2	32	30.5
AAGGTC	0			1	29	29.0
AAGT	0			1	17	17.0
AATAAC	0			1	29	29.0
AATATG	0			1	24	24.0
ACAGCG	0			1	25	25.0
ACGCG	0			1	20	20.0
AGAGGC	0			1	28	28.0
AGCCTG	0			1	24	24.0
ATCGTC	0			2	38	34.0
CCG	0			1	12	12.0
AAAAGG	0			1	28	28.0

We used linear models to determine which of the factors (size class, AT-content, GC-content flanking sequence, coding/non-coding) would efficiently predict array length (log transformed) and found only size class and AT-content were significant predictors of array length ($p<e\text{-}109$, $p<0.01$ respectively).

Overall, our results are consistent with what has been reported previously in yeast (Field and Wills 1998; Katti *et al.* 2001; Malpertuy *et al.* 2003; Lim *et al.* 2004; Karaoglu *et al.* 2005). However, some differences arise in terms of the total number of microsatellites found, the frequency of certain size classes (especially mono-, tri-, and pentanucleotides), and the detection of some rare short arrays (such as $(CCG/GGC)_4$ and several short tetra- to hexanucleotides), that appear in some studies but not in others (Katti, Ranjekar et al. 2001; Malpertuy, Dujon et al. 2003; Lim, Notley-McRobb et al. 2004). These differences

emerge mainly as a result of divergence among definition/thresholds of microsatellite search parameters and the intricacies of the search algorithm itself (Merkel and Gemmell 2008).

4.3.2. Genome wide microsatellite distribution and association with genomic context

Next, we were interested in whether or not microsatellite distribution (frequency) was uniform throughout the genome. We were also interested to determine whether the distribution of microsatellites might be associated with other genomic elements, such as autonomously replicating sequences (ARS), ARS consensus sequences, Ty-elements, long terminal repeats (LTR), tRNAs, coding regions and introns.

Overall, we found the microsatellite frequency per megabase pair increased with decreasing chromosome size (Figure 3), though the distribution along individual chromosomes appeared homogenous.

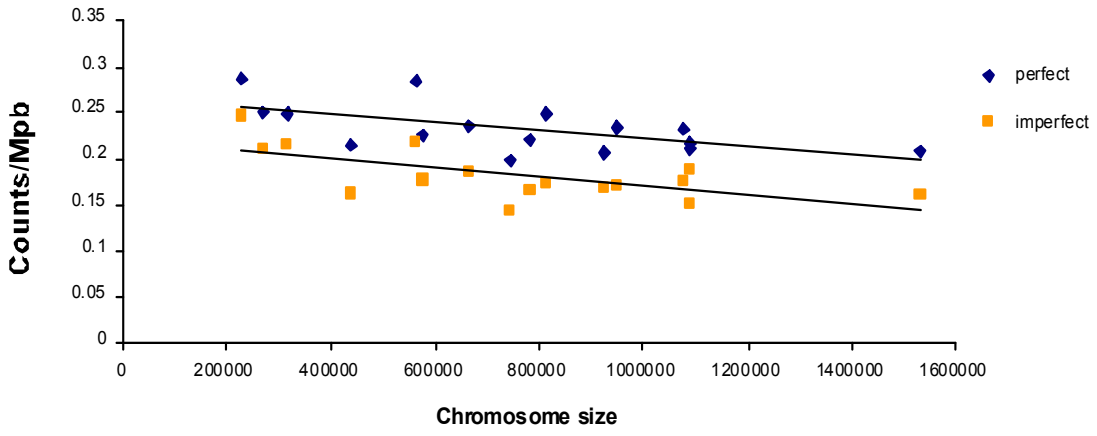


Figure 3: Microsatellite frequency for perfect and imperfect repeats across all 16 *S. cerevisiae* chromosomes. Microsatellite frequency decreases in a linear manner with increasing chromosome size for both perfect and imperfect repeats.

Nevertheless, we observed a pronounced spatial clustering between loci (Appendix Figure 3) and the intra-chromosomal density distribution showed distinguishable peaks. Based on the total number of microsatellite loci in the genome (i.e. 2732), a perfect (or imperfect) microsatellite is expected to occur on average approximately every ~ 4.5kb (~5.8kb), but in fact some regions (10kb intervals) contain up to 12 loci, 6 times the expected number. Extracting these regions of high microsatellite density (>6 repeats/10kb, genomic coverage: ~13-14%), we explored if those cluster-like regions were enriched (or depleted) for certain genomic features. Out of all features, only coding sequences were significantly underrepresented in regions of high microsatellite density (Chi-square, df = 1, $p<0.005$).

We then undertook a more thorough approach to inspect genome wide correlations between microsatellite frequency and their local genomic environment, using a generalized linear model (Table 3). We extended our analysis to incorporate several other genomic features in the analysis, such as genomic background composition, i.e. GC-content. Further, we included SNP density in the analyses to examine if microsatellites were overabundant in regions showing neutral-like evolution. Ultimately, we wanted to explore whether the distribution of microsatellites was the result of neutral processes or perhaps influenced by selective forces. First, we analysed microsatellite frequency on a small scale in 1kb windows and then repeated the analysis at a larger scale of 10kb windows.

Associations appear less abundant at a larger scale than the smaller scale, especially for individual motifs. Microsatellites are negatively correlated with SNP density and LTRs, but positively correlated with DSBs and regulatory elements. Amongst the individual motifs poly(A) arrays are most diversely affiliated, followed by $(AT)_n$ and $(AC)_n$ microsatellites. GC-content and coding sequence coverage affect almost all motifs. Notable are the scale-specific differences, such as the association between poly(A) and GC-content being reversed between smaller and larger window sizes. In other cases associations occur only in a motif specific manner, for example between nucleosomes and $(AAT)_n$ or ARS consensus sequences and telomere regions and $(AC)_n$.

Table 3. Generalized Linear Model (GLM) analysis for the association of microsatellites with genomic context features in 1kb [10kb] windows. Only results significant after Bonferroni correction for multiple hypothesis testing are shown (p< 0.003571 (All); p< 0.000325 (individual motifs)). The symbol '+' indicates a positive correlation, while '-' indicates a negative correlation.

	Predictor											
Response	**All**	**A**	**AT**	**AAT**	**AAG**	**AAC**	**ATC**	**AGC**	**AC**	**ACC**	**AGG**	*****
ARS	+		+									
ARS consensus									[+]			
CDS (coverage)	-	-	-		+		+		-			
SNP	- [-]		[-]									
GC	[+]	- [+]	[+]	-		+		+ [+]	+ [+]	+	+	+
Centromere												
DBS (*dcm1Δ*)	+ [+]	+ [+]										
DBS (*rad50S*)	+ [+]	+										
Intron		+										
LTR	- [-]											
Nucleosome				[+]								
Regulatory sites	+ [+]	+ [+]	+									
Telomere									+ [+]			
tRNA												

*other motifs

4.3.3. Scale-specific effects

In order to isolate the scale-specific influences of intra-genomic factors, such as coding sequences, ARS, DSBs, GC-content, introns, nucleosomes, regulatory elements and SNP density on the genomic distribution of microsatellites, we utilized wavelet analysis (Spencer *et al.* 2006). Under wavelet analysis a pattern (such as SNP density or microsatellite abundance along a chromosome) is transformed into a series of coefficients that describe the variation in the signal at increasing scales. The analysis of the detailed coefficients is particularly useful as it investigates changes amongst factors rather than absolute levels and is thus more robust to spurious background noise and biases in the underlying distribution than other approaches. Linear model analysis of the detailed

coefficients can be applied to estimate correlations between multiple signals at the individual level.

We performed the analysis across all 16 yeast chromosomes (see methods for details), testing the same factors as described for the GLM, but this time we evaluated their potential as significant predictors of genomic microsatellite coverage. Special focus was put on factors indicating neutral sequence evolution, such as high SNP density, and other features known to impact on these, such as coding sequence abundance, GC-content, and DSBs.

First, we examined the pairwise correlations of the wavelet coefficients for all factors at various scales using Kendall's rank test (Table 4). This test was concordant with what the GLM analysis had indicated before: a significant localized (2-8kb range) negative correlation between perfect microsatellites and coding sequences, and a significant small scale negative correlation with GC-content across all chromosomes. Strikingly, however, we discovered a significant localized positive correlation with SNP density (2kb) instead of the previously detected negative association. Additionally, there were sporadic, but non-significant, associations with other genomic features, such as DSBs and ARSs. We further observe correlation amongst factors (e.g. Appendix Figure 4), such as a consistent negative relationship between coding sequence and SNP density, which is present in all analyzed chromosomes across several scales down to the smallest level. Similarly, CDS and GC-content consistently show a very localized positive correlation.

Next, we performed a linear model analysis of the detailed wavelet coefficients which, however, due to the extreme non-normality of the data distribution (even after log-transformation) has to be treated with much caution (Table 5).

Table 4: Kendall's rank test for pairwise correlations of detailed wavelet coefficients. Shown are genomic scale (in kb) at which significant correlations occur ($p<0.001$) and type of correlation (- = negative, + = positive). After correcting for multiple hypothesis testing only correlations in bold remain significant.

Fragment size: 512kb

Chromosome	2		5		7		8		10		11		13		14		16	
Region	start	end	start	end	start	end	start	end	start	end	start	end	start	end	start	end	start	end
ARS					+32	+2												
CDS	**-2**	**-2**	**-2**	**-2**	**-2,-4**	**-2**	**-2**	**-2**	**-2**	**-2,-4,-8**	**-2**	**-2**	**-2**		**-2**	**-2**	**-2**	**-2**
DSB_dmc1					-8	+4						+16		+8				
DSB_rad50S						+4,+8					+64	-2		+16				
GC	**-2**		**-2**	**-2**	**-2**	**-2**					**-2,+64**		**-2**	**-2**		**-2**	**-2**	**-2**
Intron																-2		
Nucleosome							+16	-2,+16										
Reg. elements										+16								
SNP	**+2,+4**	**+2**	**+2**	**+2**	**+2**	**+2,+4**	**+2**	**+2**	**+2**	**+2, +4**	**+2**	**+2**	**+16**	**+2,+16**	**+2,+8**	**+2**	**+2**	**+2**

Fragment size: 256kb

Chromosome	3		6		9	
Region	start	end	start	end	start	end
ARS						
CDS		**-2**			**-2,-4**	
DSB_dmc1						
DSB_rad50S			+8			
GC			**-2**	**-2,+128**		
Intron						
Nucleosome						
Reg. elements					+2,+4	+4
SNP			**+2**	**+2**	**+2**	

Fragment size: 1024kb

Chromosome	4		12		15	
Region	start	end	start	end	start	end
ARS	-2					
CDS	**-2**		**-2,-4**	**-2,-4,-8**	**-2,-8**	**-2,-8**
DSB_dmc1	+16					
DSB_rad50S		+32				
GC	**-2**		**-2**	**-2,+32**	**-2**	**-2**
Intron						
Nucleosome						
Reg. elements						
SNP	**+2,+4**	**+2,+4**	**+2,+4**	**+2,+4**	**+2,+8**	**+2,+8**

Table 5. Simplified linear model analysis for detailed wavelet coefficients. Marginally significant results shown (-log10 p-value as determined by t-test)

		Predictor				
Chr	**Chr region**	**CDS**	**SNP**	**GC**	**Position**	**r^2**
1	start					
	end	2.13 [8] $^{-}$		3.31 [8] $^{+}$	2.15[8] $^{-}$	
2	start					
	end	2.57 [2] $^{-}$		2.59 [8] $^{+}$		
3	start					
	end					
4	start	2.52 [2] $^{-}$, 2.04 [4] $^{-}$		3.12 [2] $^{-}$		
	end	2.58 [16] $^{-}$				
5	start	2.39 [2] $^{-}$, 2.52 [16] $^{-}$		2.44 [16] $^{+}$	2.1 [16] $^{-}$	
	end	2.33[2] $^{-}$, 2.82 [16] $^{-}$		3.1 [16] $^{+}$		
6	start				2.41 [2] $^{+}$	
	end					
7	start	2.39 [4] $^{-}$		2.65 [2] $^{-}$		
	end	2.21 [2] $^{-}$, 2.23 [4] $^{-}$		2.98 [2] $^{-}$		
8	start	3.01 [2] $^{-}$				
	end	3.1 [2] $^{-}$				2.44 [8] $^{-}$
9	start					
	end					
10	start	3.6 [2] $^{-}$				
	end			2.45 [16]$^{+}$		
11	start					
	end	2.78 [8] $^{-}$		2.07 [16]$^{+}$		
12	start	3.91 [2] $^{-}$, 3.67 [4] $^{-}$				
	end	3.3 [2] $^{-}$, 3.2 [4] $^{-}$, 2.05 [8] $^{-}$				
13	start	2.02 [8] $^{-}$	2.02 [8] $^{-}$	2.8 [2] $^{-}$		
	end			2.5 [8] $^{+}$	4.17 [4] $^{+}$	
14	start	3.38 [2] $^{-}$, 2.27 [8] $^{-}$				
	end	3.76 [2] $^{-}$, 3.41 [16] $^{-}$				
15	start			3.67 [2] $^{-}$		
	end	2.31 [8] $^{-}$				
16	start					
	end			4.04 [2] $^{-}$		

+ = positive correlation, - = negative correlation, [ ] = scale in kb, r^2 = adjusted r^2, can be interpreted as the proportion of variance that is explained by the model

As before, we observe a strong localized negative effect of coding sequence, but a much more inconsistent influence of GC-content, showing a negative effect at the 2kb range and a positive effect at the 4-32kb range. SNP density and genomic position had no predictive power on microsatellite abundance while sporadic significance for DBS had to

be excluded as false positives under multiple hypothesis testing. There was no effect of any of the other features on microsatellite coverage.

The loss of SNP density as a significant effect in the linear model likely reflects the correlation between coding sequence and SNPs, thus a change of SNP density is an insignificant predictor of microsatellite abundance compared to a change in CDS coverage. Coding sequences tend to be strongly GC-rich which explains in part the relationship between GC-content and microsatellite abundance, showing negative associations at scales similar to the size of intergenic regions and positive associations at scales reflecting gene sizes in yeast. Still, the linear model predicts a significant effect of GC-content that is independent of coding sequence coverage.

Both GLM and wavelet analysis confirm the role of scale-specific effects on microsatellite distribution. Overall microsatellites tend to be subject to small scale effects rather than effects at larger scale, which is likely a result of the small intergenic regions in yeast. Microsatellite frequency is more sensitive to genomic context, i.e. shows more significant correlations, than microsatellite coverage. Finally, coding sequence absence and (to a lesser extent) GC-content of the genomic background are sufficient enough to predict local microsatellite coverage.

4.4. Discussion

The observation that even in a compact genome such as that of *S. cerevisiae* (~70% coding sequence) the majority of microsatellites (with the exception of tri- and hexanucleotide repeats) are found in non-coding regions shows that microsatellites are clearly a feature of non-coding regions. Concordantly, the genomic frequency and array composition of microsatellites reflect the strong AT-bias in non-coding regions compared to coding regions (Table 2). Together with the glaring overabundance of trinucleotides in coding regions, the only size class despite hexanucleotides not causing frameshift mutations, everything points towards a neutral model of evolution for microsatellites, a

finding strongly supported by previous studies (Goldstein and Schlötterer 1999; Toth *et al.* 2000; Katti *et al.* 2001; Dieringer and Schlotterer 2003). However, as the strong influences of coding/ non-coding regions on microsatellite type and frequency show, microsatellites are genomic entities, though their implications on genome dynamics and *vice versa* are only slowly being revealed.

4.4.1. Genomic elements

LTR retrotransposons

In regards to the origin of microsatellites, LTR retrotransposons are certainly not a causal agent in yeast. The observed negative association between retrotransposons and microsatellite frequency in yeast (Table 3) is much more similar to the pattern observed in plants rather than that observed in insects or mammals. Plant LTR retrotransposons have no poly(A) tails or proto-microsatellite-like structures from which a mature microsatellite could possibly emerge. Further, large regions of plant genomes filled with LTR retrotransposons are depleted of microsatellites, which has been explained with the recent expansions of these regions through retrotransposon propagation and a delayed population of these regions with microsatellites (Morgante *et al.* 2002). In contrast, in insects and mammals the pattern is reversed, with strong associations between retrotransposons and microsatellites (such as *Alu* and L1 elements in human, and *mini-me* elements in dipterans, respectively (Duffy *et al.* 1996; Nadir *et al.* 1996; Wilder and Hollocher 2001)).

ARS

Our results show a local (1kb) association of $(AT)_n$ with ARS elements, which, however, have an elevated AT-content that has been related to strand separation properties. Further, $(AT)_n$ have very low Z-DNA structure-forming potential far behind for example $(AC)_n$ or $(GC)_n$ (Herbert and Rich 1996). Hence, ARS regions are more likely to act as seeding

grounds for microsatellites rather than being related to microsatellites in a functional way.

Amazingly, more than a decade ago, Valle et al. (Valle 1993) had reported that on yeast chromosome 3 the nine most dense clusters of $(AT)_n$ microsatellites were enriched in ARS consensus sequences. However, despite ARS consensus sequences being essential for functional replication origins in yeast, since they bind one of the replication initiation proteins, the Origin Recognition Complex (ORC), not all ARS consensus sequences function as replication origins (Newlon and Theis 2002). A fair proportion resemble integral components of the telomere X-element (Louis *et al.* 1994). In fact, we observe an overabundance of $(AC)_n$ near ARS consensus sequences when investigating ARS consensus sequences only. Further, since $TG_{1\text{-}3}$ (or $AC_{1\text{-}3}$) are well characterized terminal telomere repeats in *S. cerevisiae* as well as being occasionally found between the X- and Y elements (Louis *et al.* 1994), and we observe an association of $(AC)_n$ with telomere regions, we attribute the association of $(AC)_n$ repeats with ARS consensus sequences to their abundance in telomeric regions, rather than functional properties of replication origins.

Nucleosomes & regulatory elements

We observed a significant overrepresentation of poly(A/T) in regulatory regions, which has been implicated before in transcription regulation via nucleosome depletion even in a length dependent manner (Iyer and Struhl 1995). Given this exclusionary relationship between poly(A/T) and nucleosomes we expected to detect a negative relationship between these features but we did not. There are several reasons that could explain why we were not able to detect a negative association between poly(A/T) and nucleosomes: First, the association may be at a much smaller scale than we investigated, e.g. hundreds of base pairs instead of thousands since the DNA stretch wrapped engaged within a single nucleosome is only about 147bp (Segal *et al.* 2006). Second, while their rigidity does not promote it, nucleosomes can form in poly(A/T) rich regions (Losa *et al.* 1990). Third, these elements themselves could bind transcription factors rather than having to mediate their effects indirectly via nucleosome exclusion. However the only protein identified so

far that binds poly(A/T) is datin (Moreira *et al.* 1998), so if these elements were important binding sites for transcription factors we might have expected to have found more such proteins by now. We believe, that employing an inappropriate scale is the most likely explanation for our inability to detect the previously observed negative association between poly(A/T) and nucleosomes, and, for the future, we suggest that the analysis be repeated at a smaller scale (~100bp)..

However, we find a larger scale (10kb) positive association between $(AAT)_n$ repeats and nucleosomes which could be explained with the thermodynamically preferred positioning of nucleosomes caused by periodically oscillating AA/AT/TT dinucleotides (Satchwell *et al.* 1986; Widom 2001). This dinucleotide periodicity promotes a sharp bending of the DNA molecule at every helical repeat (~10bp) and facilitates the wrapping around the nucleosome. The $(AAT)_n$ repeat approximates this periodicity, i.e. AA|TA|AT|AA|TA|AT. An overrepresentation of DNA sequences with high nucleosome affinity in certain regions of the genome may attract nucleosomes to preferentially reside in these regions, making those regions subject to intense epigenetic regulations. Segal et al (2006) have shown that overall the yeast and chicken genomes is enriched in sequences that show high nucleosome affinity.

There is general support for a genome wide association of microsatellites with regulatory regions, but no other motif-specific association could be identified. The highly compact architecture of the yeast genome means that most regulatory elements are located within small intergenic regions, and the genome wide association of microsatellites with regulatory regions might be an artifact, arising as a consequence of this architecture. Generally, transcription factor binding sites are constituted from larger rather than shorter motifs. However, motifs of larger unit sizes occur less frequently than shorter unit sizes, thus an association between such motifs and regulatory regions could likely remain undetected. In keeping with this prediction we did observe a weak but insignificant correlation for motifs other than the 10 most abundant groups ($p<0.008$, Bonferroni $\alpha= 0.000325$).

Double Strand Breaks

Despite the five-fold overabundance and somewhat more even distribution of *dmc1*Δ derived DSBs compared to those derived from *rad50S* mutants (Buhler *et al.* 2007), we observe no differences in their association with microsatellites. In both knockout backgrounds, we find positive associations for poly(A) and the total genomic microsatellite set (Tables 3, 4). Recent studies have already shown a significant association of *rad50S* derived "hotspot" DSBs especially with mono- (≥14 copies) but also dinucleotides (>6 copies) in yeast (Bagshaw *et al.* 2008). Although we observed the latter association involving dinucleotides repeats, we found it was not significant after our GLM analysis ($p<0.0007$; after Bonferroni $\alpha = 0.0003$).

The mechanistic basis underlying these correlations are however uncertain. The observation that high frequencies of poly(A) are also present outside "hotspots" and can be absent from functional "hotspots" (Bagshaw *et al.* 2008) indicates that these sequences function interactively rather than as solitary entities. Several microsatellites are known to bind transcription factors and alter DNA structure, which in turn has been reported to be of functional significance for recombination hotspots (Petes 2001; Nishant and Rao 2006). Poly(A) arrays show only very limited protein binding effects, but rather act as nucleosome depleting sequences (see above) which has been shown to stimulate recombination by means of DNase I hypersensitivity in some, but not all, hotspots (Petes 2001). In fact, the deletion of a poly(A) at the *ARG4* locus in yeast resulted in dramatically reduced recombination activity (Schultes and Szostak 1991). An alternative view, but not necessarily conflicting position, is that microsatellite abundance might be a result of DSB repair activity and recombination contributes actively to microsatellite mutation (see *Chapter 1*, 'Mutation mechanism(s)'). In any case, the association with *dcm1*Δ derived DSB, which are about five times more common than *rad50*Δ DSB (Buhler *et al.* 2007), implies that such an association involves a much larger fraction of the genome than previously thought.

4.4.2. Microsatellite evolution

Our results show that evolutionary constraints imposed by coding regions act strongly and precisely on microsatellites, as the small scale negative relationship between microsatellites and coding regions coincides with the small size of intergenic regions in yeast (more than half of all intergenic regions are smaller than 1kb) (Goffeau *et al.* 1996). However, such constraints do not appear to apply against length polymorphism *per se* (or at least to a much lesser extent), since there is no significant difference observed in array length for microsatellites in different genomic fractions, not even within individual size classes. Alternatively, microsatellite turn-over might be much higher in non-coding regions than coding regions, which appears reasonable considering that functionality has been shown for several homopolymers. For example proline and glutamine stretches act as transcriptional activation domains in transcription factors (Gerber *et al.* 1994)

The cryptic relationship between microsatellites and genomic background sequence is particularly noteworthy. On a localized genomic scale (1-2kb) GC-content is negatively associated with microsatellite abundance, whereas on a larger scale (8-16kb) we observe the reverse trend. The former case (small scale association) is plausibly explained by the AT-richness of non-coding regions. But with regards to the independent effect (i.e. independent of other factors) seen in the linear model analysis, it requires an additional explanation: Despite retrotransposon initiation, microsatellites originate through small insertion/ deletion events, missed by the mismatch repair machinery, that eventually establish a number of repeat copies sufficient enough to undergo slippage. However, mismatch repair (MMR) varies throughout the genome and has, at least in yeast and humans, a bias that leads to increased GC-content, from which it has been suggested that regions of high GC-content might represent regions with relatively effective MMR (Brown and Jiricny 1989; Birdsell 2002). Therefore, the negative association between microsatellites and GC-content could be due to the preferential origin of microsatellites in AT-rich regions due to less efficient mismatch repair. That this might even occur at a very small scale is supported by our finding that microsatellite array composition in non-coding regions is more distinct from its flanking sequence composition, generally more AT-enriched than microsatellites found in coding regions, where mismatch repair acts on

entire transcription units (Figure 1). Nevertheless, we are unable to derive a plausible explanation for the larger positive scale association (8-16kb) with GC-content since the average ORF size in yeast is usually smaller (~ 6kb).

That SNP density has a negative effect on microsatellite frequency (Table 3) supports the idea of microsatellite death due to point mutations (Taylor *et al.* 1999). Initially, point mutations accumulate in the repeat array over time and stabilize the array due to imperfect copies decreasing the possibility of strand misalignment and subsequent polymerase slippage. Eventually though such point mutations degrade the array so much that it becomes unrecognizable against the genomic background. Arguably, a point mutation splits a perfect array into several shorter arrays, which should result in increased frequency of (perfect) microsatellites in regions with high SNP density. Nevertheless, since the majority of microsatellites in a genome are short and substitutions within the array are commonly polar, i.e. occur preferentially towards the array end (Brohede and Ellegren 1999), the resulting segments are likely to fall below the minimum detection threshold and are not detected as microsatellites. Consequently, this would incur an increase of imperfect microsatellites in regions of high SNP density above genome wide average (which may also be an interesting hypothesis to test in the future).

Furthermore, point mutations have been implicated in the 'balanced model' of microsatellite evolution (Kruglyak *et al.* 1998; Kruglyak *et al.* 2000). Under the balanced model the observed distribution of microsatellite length in the genome is a result of the balance between slippage and point mutation and this model may explain the observed upper length limit for genomic microsatellites (Kruglyak *et al.* 1998; Kruglyak *et al.* 2000). Kruglyak *et al*.. (1998, 2000) proposed the model based on the observed length distribution in 1Mb long segments in human, mouse, yeast and fruitfly and later for the entire yeast genome. Our results support Kruglyak's model at a local scale in addition to its common application at a genomic scale. Considering that following the predictions of the model microsatellites would preferentially mutate in regions with low SNP density, this would greatly facilitate association studies.

4.5. Summary

Microsatellites are predominantly neutrally evolving sequences with a genomic distribution that is mediated by local sequence composition, selective constraints, and reflects DNA metabolic processes, such as repair, strand slippage and point mutation. High SNP density is a negative predictor for, at least, perfect microsatellite frequency due to the stabilizing/ degrading effect of substitutions on microsatellites. However, certain genomic features, such as ARS or LTRs, additionally affect the pattern of microsatellite distribution by facilitating the emergence or depletion of microsatellites in certain regions, respectively. In addition, microsatellites have associations with genomic elements such as meiotic double strand breaks, regulatory sites, and nucleosomes, which supports a functional role for some microsatellites.

Chapter 5

Conservation of Microsatellites in the Yeast Genome

5.1. Introduction

Microsatellites are a common element of most genomes. An extraordinarily high level of length polymorphism has made these short (1-6bp long) tandem repeats versatile genetic markers for a variety of applications, such as linkage analysis, population genetics, pedigree analysis and DNA forensics (Goldstein and Schlotterer 1999). Following on from their role as genetic markers, microsatellites are anticipated to evolve neutrally without any particular biological purpose. However, several studies have indicated that microsatellites have the potential to diverge from such expectations. First, the conservation of some loci over long evolutionary time scales may imply some level of functional or evolutionary constraint (FitzSimmons *et al.* 1995; Coote and Bruford 1996; Primmer and Ellegren 1998). Second, the association of microsatellites with other genomic features, such as recombination hotspots, may designate some structural importance (Bagshaw *et al.* 2008, see *Chapter 4*). Third, microsatellites can influence gene expression via their frequent and reversible length alterations caused by DNA strand slippage (Kashi and King 2006).

Phenotypic effects generated by variable microsatellites are reported by an increasing number of studies. For example, there are more than 30 known human neurodegenerative disorders that are caused by excessive expansion of trinucleotide repeats (Gatchel and Zoghbi 2005; Pearson *et al.* 2005; Mirkin 2007). Located within virtually every part of a gene (i.e. introns, exons, UTRs) trinucleotide loci can cause the inhibition of transcriptional elongation, DNA hypermethylation and/or lead to altered mRNA and protein functions (Gatchel and Zoghbi 2005; Pearson *et al.* 2005). In promoter regions microsatellite variability can affect transcription factor binding efficiency or interrupt the

highly sensitive interactions between transcription factors resulting in altered gene expression (Kashi and King 2006). The resulting phenotypic effects are as diverse as the underlying mechanisms that these promoters mediate, which span circadian cycle and temperature adaptation in fruit flies, social behaviour in voles, scull morphology in dogs and others (Sawyer *et al.* 1997; Fondon and Garner 2004; Hammock and Young 2005); for a review see (Kashi and King 2006)).The number of studies in which microsatellites are implicated increases further if one considers the large number of microsatellite loci used in QTL mapping studies, whereby a defined microsatellite allele is linked to a certain phenotype, but it might also be the cause of it.

More importantly, there is evidence that the variability induced by microsatellite length polymorphism can provide an evolutionary advantage. A classic example are the so called contingency genes found in many disease causing bacteria. Here the expansion or contraction of a microsatellite results in the disruption of protein synthesis and/or, following reoccurring mutations the (re)gain of protein function (for a recent review see Moxon *et al.* 2006). The affected genes mostly encode cell surface molecules that determine the pathogen's adherence to the host or, alternatively, it's susceptibility to the host's immune attack (Moxon *et al.* 2006). Consequently, the frequent on-off switching of those genes generates a functional diversity that allows for rapid adaptation of the pathogen to its host environment (Moxon *et al.* 2006)

It is unlikely that such functionality is exhibited by a large fraction of genomic microsatellite loci, since it is difficult to imagine how selection could sustain such a large number of phenotypes. Nevertheless, there are only a few studies that have taken a genomic approach to confirm or withdraw such assumption. One of these, a recent genome-wide approach undertaken by Verstrepen et al (2005) to investigate the role(s) of intragenic tandem repeats in *S. cerevisiae,* revealed that variable tandem repeats with large repeat units (>40nt), i.e. minisatellites (tandem repeats with unit size >10bp – 100bp (Jeffreys *et al.* 1994)), were significantly overrepresented in genes encoding yeast cell wall and cell surface proteins. Since many fungi are opportunistic pathogens, this variability has been implicated to cause a fitness benefit similar to that seen for

contingency loci in pathogenic bacteria (Verstrepen *et al.* 2005). In fact, similar had previously been suggested for a group of adhesins (cell wall glycoproteins), namely *FLO* (flocculation) genes in *S. cerevisae* and *ALS* (agglutinin-like sequences) genes in *Candida albicans* (Zhang *et al.* 2003; Verstrepen *et al.* 2004). However, no specific gene function, was found to be associated with intragenic microsatellite polymorphism (Verstrepen *et al.* 2005). Nevertheless this might be due to the limited scope of the study towards microsatellites: only tri-to hexanucleotides were included and genotyping was carried out across six strains.

Other previous genomic investigations in *S. cerevisiae* report trinucleotide repeats to be particularly strongly overrepresented in genes encoding products found in the nucleus (specifically transcription factors) and proteins involved in signal transduction, i.e. cellular regulation (Alba *et al.* 1999; Young *et al.* 2000). The same observation has been made for lineage specific transcription factors across 13 individual species spanning the *Hemiascomycete* phylum (Malpertuy *et al.* 2003). However, there has been no polymorphism data to support any influence of microsatellite variability on these genes.

Here we utilize recently available data on 40 complete sequenced *S. cerevisiae* strains (*Saccharomyces Genome Resequencing Project*, http://www.sanger.ac.uk/Teams/Team71 /durbin/ sgrp/browser.shtml) to describe the conservation and variability of microsatellite loci in the yeast genome. Under a neutral model, we expect microsatellites to be conserved at random; with those microsatellites conserved representing a subpopulation of the present genome wide set. In contrast, any divergence from the background pattern of microsatellite distribution, i.e. our null expectation, may implicate functional roles for those microsatellites. We further analyze the pattern of overall microsatellites polymorphism, in order to identify determinants of these patterns and allow for a better description of microsatellite evolutionary dynamics. Particularly, we investigated biases in locus conservation (and polymorphism) (i) along individual chromosomes, (ii) between loci related factors and genomic position, and (iii) evaluated protein function associated with microsatellites located in genes and promoters.

5.2. Methods

Extraction of conserved microsatellites

First, we searched for perfect (genomic) microsatellites in the *S.cerevisiae* genome sequence (SGD, ftp://genome-ftp.stanford.edu/pub/yeast/; ref |NC_001133 – NC_001148|; accessed November 2007) using the SciRoko microsatellite analysis tool (Kofler *et al.* 2007) employing a minimum length threshold of 12bp and 4 repeat copies for mono-to trinucleotides and tetra- to hexanucleotides, respectively (see also *Chapter 4*). Next, we identified conserved microsatellites in *S. cerevisiae*, by performing electronic PCR (i.e. re-PCR version 3.0 (Schuler 1998)) on 40 fully sequenced *S. cerevisiae* strains (SGRP, ftp://ftp.sanger.ac.uk/pub/dmc/yeast/latest, November 2007). For our purpose 'electronic PCR' (E-PCR), had several advantages over a BLAST or BLAT search, such as close matching of both flanking sequences (i.e. the PCR primers) in the correct order, orientation, and spacing (Schuler 1998)). We used 2732 primer pairs of 30bp length derived from the flanking sequences of the previously detected perfect microsatellite loci to perform the search. The E-PCR parameters we used were as follows: minimum seed length = 12, maximal 2 mismatches, no gaps allowed (other parameter settings did not reveal significantly more unique matches (data not shown)).
Some loci were present in multiple copies within the reference sequence, due to localization in segmental duplications, gene families, protein domains or telomeric repeats. To select only unique hits, we filtered all results for matches that were found within a range of +/-70bp to the reference locus, and that occurred 40 times or less i.e. a maximum of one time per strain. Finally, we defined a locus as conserved when it was found at least in the reference sequence and two other strains.

File handling and data manipulation was achieved through in house Perl scripts. For filtering and extracting conserved microsatellites, we executed customized queries on a purpose build database implementing MySQL and Bioperl modules. All scripts are available on request.

Genome wide microsatellite distribution and statistical analysis

We separately calculated frequencies for genomic, conserved and conserved variable microsatellites along each chromosome in 10 kb sliding window intervals (100bp overlap). Next we investigated putative spatial associations between microsatellites and other genomic features (1kb windows) (i.e. coding sequences, SNP, long transposable repeats (LTR), autonomously replicating sequences (ARS), double-strand breaks (DSB), introns, tRNAs, GC-content and regulatory sites, see *Chapter 4* for references). Each element was assigned into a 1kb interval based on its start position and whether the majority of nucleotides (>50%) were located in the interval, otherwise the element was assigned to the next interval (for coding sequences and SNP, coverage and density was used instead). As the data was strongly skewed towards zero and experienced overdispersion, we implemented a quasi-Poisson Generalized Linear Model (GLM) for the analysis. We also utilized wavelet transformation to further explore scale specific effects, but used nucleotide coverage instead of simple counts as a measure. Specifically, we tested pairwise ranked correlations of the detailed wavelet coefficients for significance (after (Spencer *et al.* 2006), for a more detailed description of the analysis see *Chapter 4*).

Distribution bias for conserved and variable conserved (polymorphic) microsatellites in different genomic regions/ elements were analyzed by testing for equal proportions. Because the test is based on the Pearson χ^2 statistic, we tested only regions/elements with at least 8 microsatellite occurrences to exclude misleading probabilities (Quinn and Keough 2002). All statistical analysis was carried out using R statistical software, version 2.5.0 (http://cran.r-project.org/). Polymorphism Empirical Cumulative Distribution Functions (ECDF) were calculated using Matlab® version 7.0. Coverage of genomic intervals was estimated using the Galaxy web-server (Giardine *et al.* 2005) (http://g2.trac.bx.psu.edu/).

Microsatellites and functional associations

Nucleotide sequences of coding microsatellites were translated into amino acids (in the appropriated reading frame) using the ExPasy translate tool (http://www.expasy.ch/tools/dna.html). To identify if the genes containing microsatellites (or genes associated with promoter regions containing microsatellites) constituted a functionally distinct group of genes, we employed the Gene Ontology tool 'GO-term finder' after Boyle *et al.* (2004) as it is implemented on the SGD homepage (http://db.yeastgenome.org/cgibin/GO/goTermFinder.pl; March 2008). The tool identifies significantly shared GO-terms (a controlled vocabulary to describe gene and gene product attributes) amongst a selected set of genes by comparing it to a background set of genes. Significance of a 'GO-term cluster' is calculated by p-values based on the hypergeometric distribution using Bonferroni's correction for multiple hypothesis testing. The hypergeometric distribution is a probability distribution that allows the estimation of the number of successful draws in a finite population without replacement. The number of genes that are annotated with a specific GO-term within all yeast ORFs (7291) function as reference, so that GO-clusters are unbiased by the enrichment of the same GO-term in the background set. Specifically, we used GO-term finder to test for functional differences in gene sets containing (i) genomic microsatellites, (ii) all conserved microsatellites and (iii) polymorphic microsatellites only.

5.3. Results

5.3.1. Distribution of conserved microsatellites

Out of 2372 microsatellite loci originally detected in the reference sequence, we identified 893 unique loci as conserved (excluding 82 paralogous primer pairs and 64 that were not recognized at all through E-PCR). Inspecting the distribution across individual chromosomes shows that these loci occur in distinct blocks throughout the genome

covering approximately 30- 40% of the total genome (Figure 1). Amongst those, 268 loci (32%) are polymorphic and appear rather randomly distributed.

With respect to their immediate genomic environment (1kb region), we find that, similar to genomic microsatellites (see *Chapter 4*), the frequency of conserved microsatellites is negatively associated with coding sequences (CDS), Single Nucleotide Polymorphism (SNP) and Long Transposable Repeats (LTR) ($p<0.01$, quasi-Poisson GLM), but have no significant associations with any other genomic element/feature (such as autonomous replicating sequence, meiotic double-strand breaks, introns, nucleosome position, regulatory sites, tRNA, GC-content). Further, wavelet analysis (see methods and *Chapter 4*) shows that significant influences on microsatellites coverage are mostly restricted to small scales (2kb), reflecting the evolutionary dynamics of non-coding regions. Here, we also observe a small scale negative correlation with CDS, a small scale positive correlation with SNP density (likely the indicator of non-coding regions), and a mixed correlation (small scale negative/ larger scale positive) with genomic GC-content ($p<0.01$, Kendall rank test of detailed wavelet coefficients).

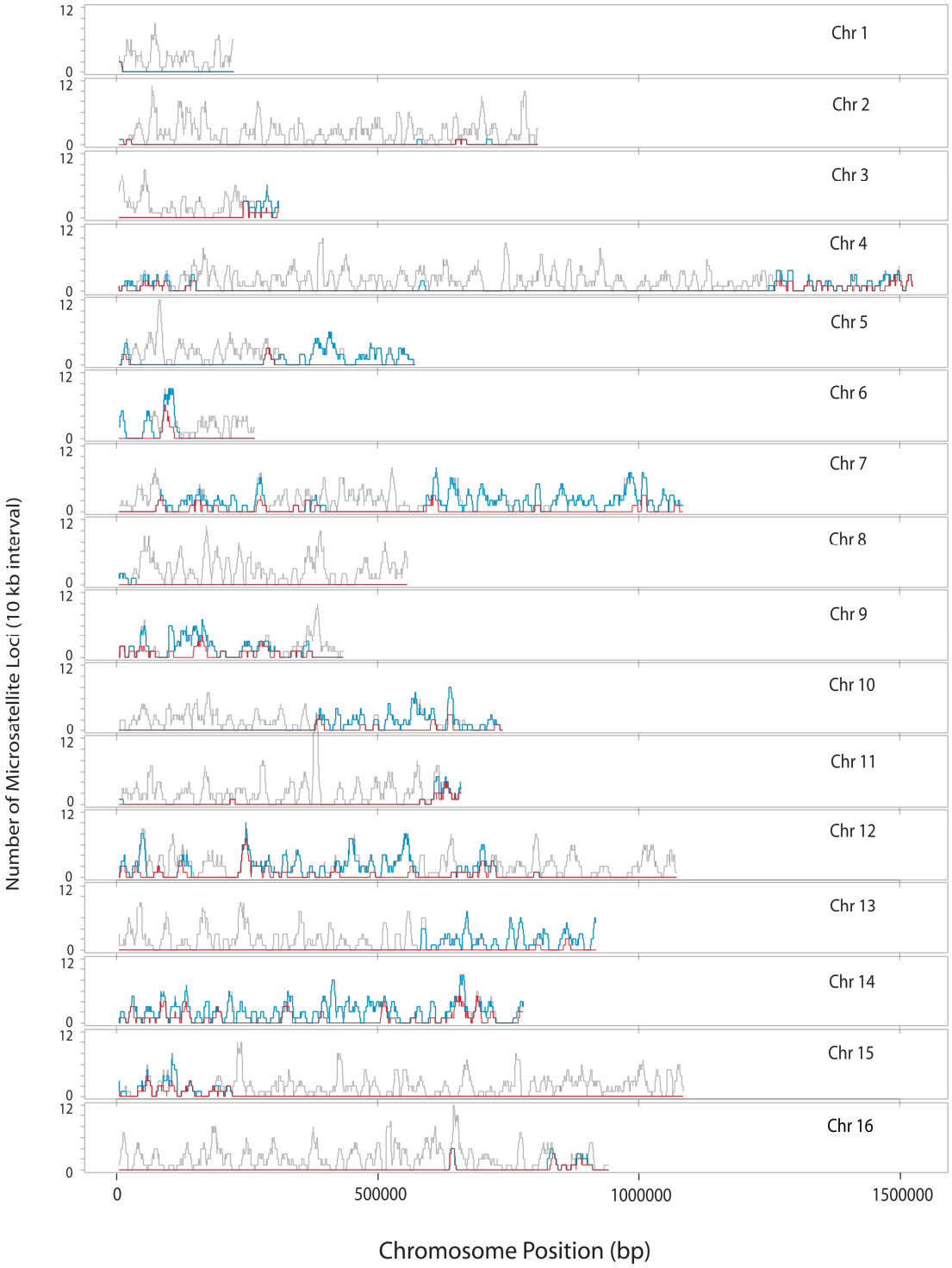


Figure 1. Distribution of microsatellites across individual yeast chromosomes (genomic microsatellites are coloured in black, conserved microsatellites in blue, polymorphic conserved microsatellites in red).

Genomic fractions

Locus conservation is significantly different within genomic fractions/ elements (6-sample test for equal proportions, Chi-square, p< 8.665e-10) (Table 1). However, polymorphism occurs at similar rates amongst genomic fractions/ elements (4-sample test for equal proportions, Chi-square, p = 0.22). Coding sequences show the lowest amount of conserved microsatellite loci, whereas microsatellites in non-annotated (i.e. non-coding) genome regions are proportionally most conserved. The lowest number of polymorphic microsatellites was detected in promoter regions, followed by coding regions, which suggests a strong negative selection against microsatellite length polymorphism in these regions. Microsatellites in genomic regions lacking annotation (not annotated) were most variable. A few variable conserved microsatellites are also found in telomere regions and autonomously replicating sequences.

Table 1: Microsatellites in different genomic fractions/ elements

Genomic element	Non-conserved microsatellites	Conserved microsatellites	Conserved microsatellites (non-variable)	Conserved microsatellites (variable)
CDS	764	285 **	202	83[1]
intron	17	9 **	6	3
promoter	723	289 **	207	82[1]
pseudogene	1	1	-	1
retrotransposon	3	2	2	-
ARS	38	19 **	11	8[1]
telomere	16	11**	7	4
ncRNA	7	-		-
rRNA	2	3	3	
snoRNA	2	-		-
tRNA	1	-	-	-
not annotated	465	330 **	217	113[1]

CDS = coding sequences, ARS = autonomously replicating sequences;

** 6-sample test of equal proportions (p <0.001), [1] 4-sample test of equal proportions

Repeat unit size

Representing the bulk of microsatellite loci, di- and mononucleotides are most conserved followed by trinucleotides (Table 2), which show only low conservation.

Table 2: Distribution of microsatellites with different motif lengths

Motif length (bp)	Genomic loci	Conserved loci (% genomic loci)	Polymorphic loci (% conserved loci)
1	1120	401 (35.8)	122 (30.4)
2	449	162 (36.1)	58 (35.8)
3	1020	281 (27.6)	88 (31.2)
4	63	20 (31.7)	9 (45.0)
5	29	7 (24.1)	2 (28.6)
6	51	15 (29.4)	3 (20.0)

Polymorphism occurs in a similar fashion though dinucleotides reveal a slightly higher fraction of polymorphic loci than mono-and trinucleotides. Exceptional high polymorphism occurs in tetranucleotides, although this might be distorted by their rarity. Pentanucleotides are the least conserved, while hexanucleotides are least polymorphic.

That microsatellites with different repeat motif lengths show different conservation/ polymorphism, is partly a result of their genomic distribution. The frequency of microsatellites of different size classes differ amongst genomic fractions, with a characteristic overabundance of trinucleotide repeats in coding regions (Messier *et al.* 1996) and an exponential decrease of microsatellites with increasing repeat unit size in promoter and other non-coding/ not-annotated regions (Figure 2). Almost all conserved polymorphic loci we observe in coding regions are trinucleotides. When comparing non-coding regions, we find that, despite very similar overall trends, the distribution of size classes differ slightly, probably indicating influences on promoter regions caused by their proximity to coding sequences. For example, the frequency of mononucleotides is slightly lower in promoter regions than in other non-coding regions, whereas the frequency of dinucleotides is slightly higher.

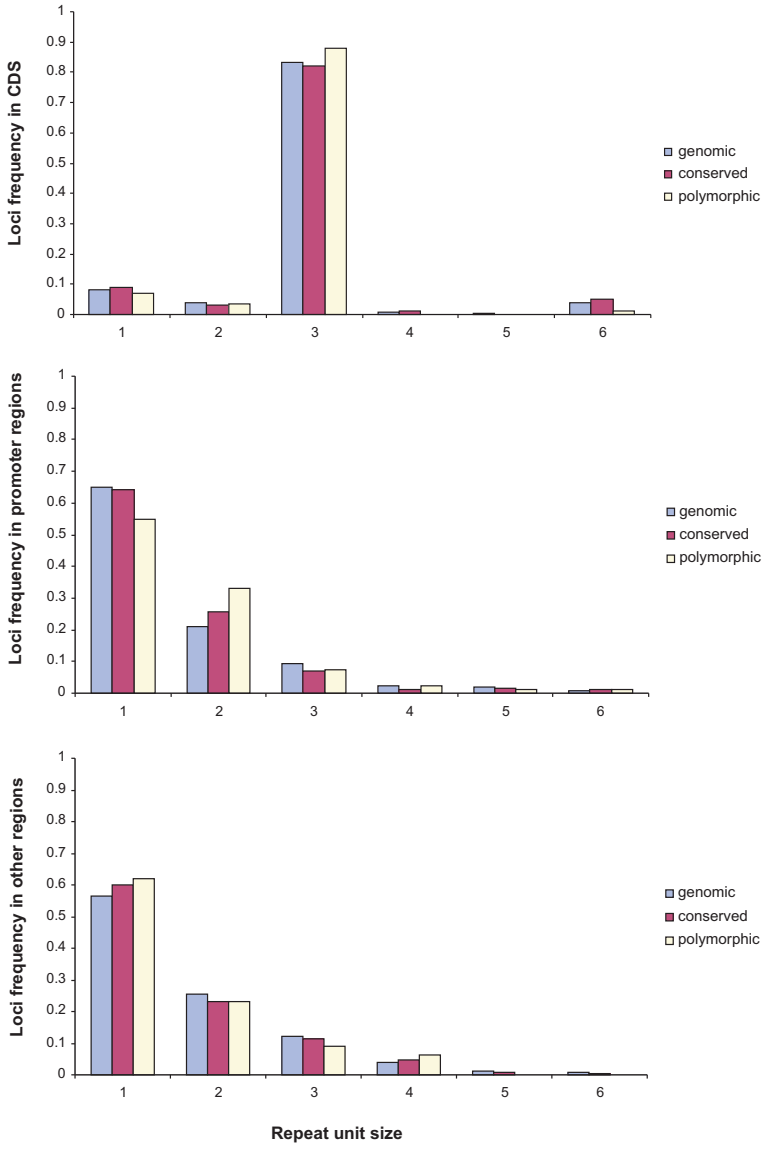


Figure 2. Distribution of microsatellite size classes in coding sequences (CDS), promoter regions and other (non-coding) regions

Motif type

The majority of motifs experience at least some level of conservation, but about a third of all motifs are lost very quickly over time. Motif conservation is frequency dependent and does not differ between motifs in coding and motifs in non-coding regions (Figure 3).

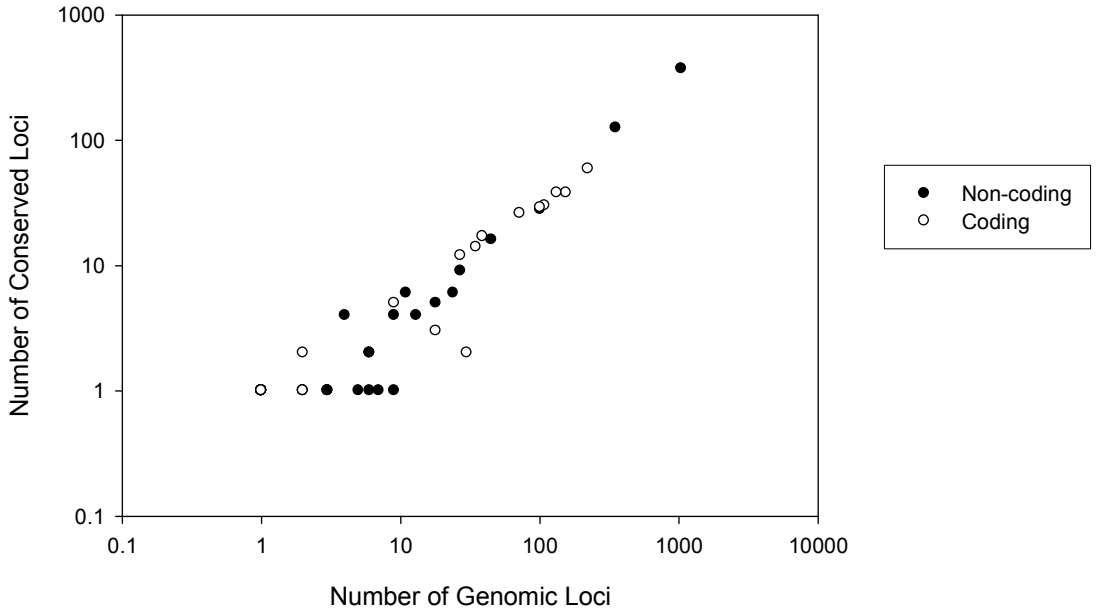


Figure 3. Conservation of motif types in coding and non-coding regions

Following a pattern similar to that observed for the total population of microsatellites found in the genome, those loci found to persist across strains are predominantly AT-rich motifs, such as poly(A) and AT_n. Most other motifs are rare and motif distribution differs significantly amongst coding and non-coding regions with a strong dominance of trinucleotide repeats in coding regions. In non-coding regions the most common motifs for conserved loci are A_n (371), AT_n (125), AAT_n (28) and AC_n (16), whereas in coding regions these are AAG_n (59), AAC_n (38), ATC_n (38) and AAT_n (30) (see Appendix).

The overall conservation of microsatellite loci is low and polymorphism is rare: only about 7.3% are conserved in more than 10 strains and ~2.4% of all conserved loci have more than 3 alleles. The most conserved and polymorphic loci (strains, alleles) are A_n (35, 7), AT_n (37, 6), AGC_n (31, 7), AAG_n (35, 5), whereas the most polymorphic loci (alleles/strains) are A_n (4/4, 4/4, 3/3, 3/3) (see Appendix). All of which indicates frequency dependent polymorphism.

Array length

Microsatellite frequency decreases exponentially with increasing array length. There is no significant difference in either average array length between genomic (16.3bp ± 6.3) and conserved loci (~16.1bp ± 5.5) or the length distributions themselves (Figure 4).

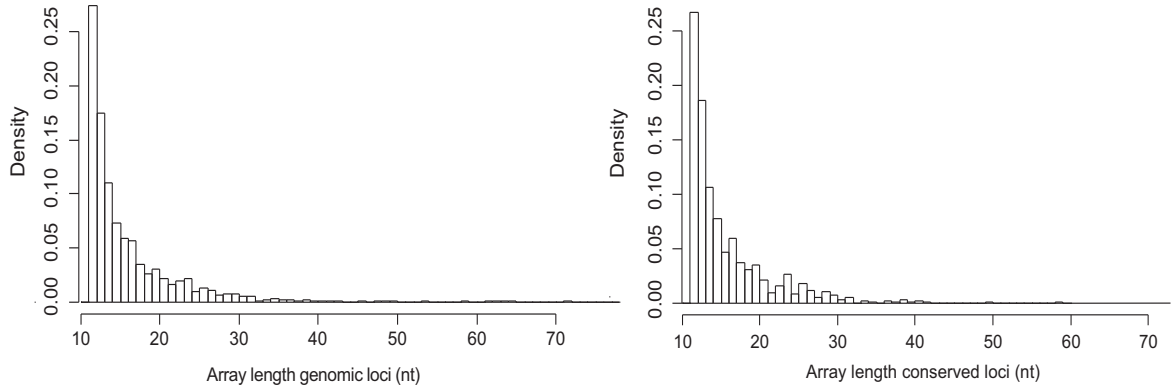


Figure 4. Length distribution of genomic and conserved microsatellites

5.3.2. Polymorphism & mutational dynamics

Microsatellite polymorphism itself, i.e. number of alleles per number of strains (K\N), is non-uniformly distributed amongst loci (Figure 5). Surprisingly, however, the distribution across different genomic fractions such as coding sequences (CDS), promoter regions and non-annotated regions is very similar, although we expect evolutionary forces, such as selection, to act on each region differentially. Compared with the fraction-specific polymorphism distribution, the size-class-specific polymorphism distributions show a much larger divergence. Hence, microsatellite polymorphism appears more likely influenced by individual repeat unit sizes rather than by genomic environment. Despite the lack of statistical significance, it seems that the underlying distributions in CDS and promoter regions are more similar to each other than they are to the cumulative

distribution of polymorphism in not annotated regions. Similarly, the distribution for dinucleotides appears most distinctive from those for mono- and trinucleotides.

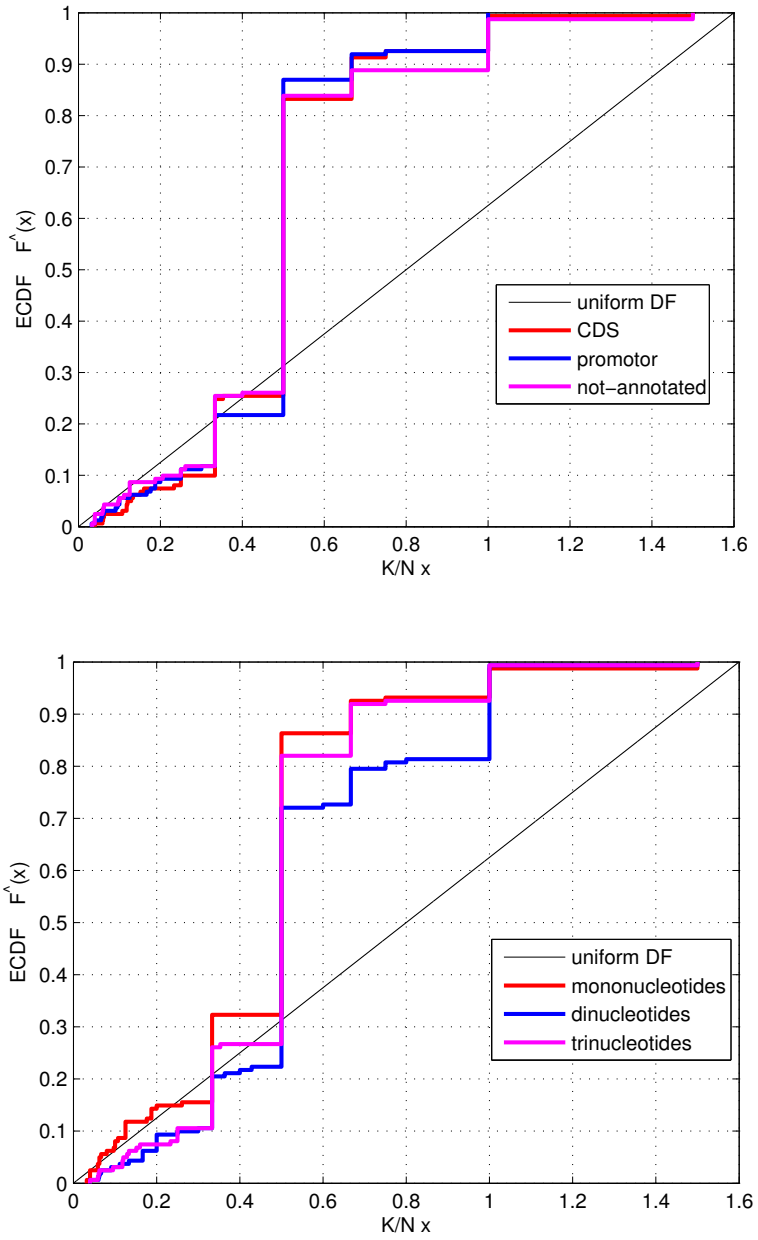


Figure 5. Nonparametric estimations of the fraction-specific polymorphism distribution (top), and the size-class-specific polymorphism distribution (bottom) for mono-, di- and trinucleotide repeats. K/N = number of alleles per locus/ number of strains, ECDF = Empirical Cumulative Distribution Function.

Array length

Polymorphism correlates positively with array length - an observation made by several other studies and probably the most dominant influence on microsatellite mutation rate known to date (Wierdl *et al.* 1997; Brinkmann *et al.* 1998; Primmer *et al.* 1998; Schug *et al.* 1998; Ellegren 2000; Kelkar *et al.* 2008).

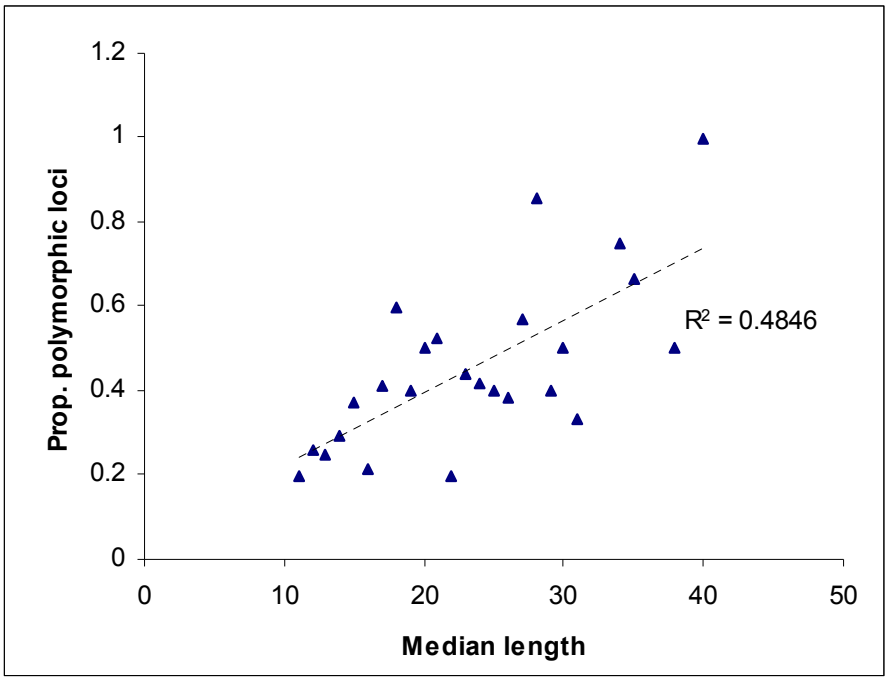


Figure 6. Median array length is positively correlated with microsatellite polymorphism (min. median length >10nt, prop. polymorphic loci >0, single occurrences per array length were excluded from the analysis)

The correlation is explained by DNA strand slippage which is thought to be the mutation mechanism responsible for microsatellite length polymorphism (Levinson and Gutman 1987b). As the DNA strands separate, a secondary structure/ loop is formed and followed by strand misalignment due to the repetitive nature of microsatellites, polymerase slippage occurs. A longer repeat array allows more opportunities for loop formation/ misalignment, and hence creates more frequent mutations than shorter arrays.

5.3.3. Functional associations

In total, we detected perfect microsatellites within 798 genes and 934 promoter regions. Out of these 233 genes and 269 promoter regions contained conserved microsatellites, including 77 genes and 79 promoter regions with polymorphic microsatellites.

After filtering for dubious/ uncharacterised ORFs, we found 223 homopolymer and 10 heteropolymer amino acid stretches to be encoded by microsatellites (see Table 3).

Table 3. Amino acid stretches encoded by perfect conserved microsatellites

Amino acid	Monomorphic loci	Polymorphic loci	Repeat size
Ala	5	5	
Cys	0	0	
Asp	12	12	
Glu	28[#]	15	
Phe	3	6	
Gly	5[++]	-	
His	4	-	
Ile	2	-	
Lys	8[++]	4[+]	
Leu	3	1	3
Met	1	-	
Asn	26	7	
Pro	4	2	
Gln	30	11	
Arg	1	1	
Ser	16	5	
Thr	1	1	
Val	1	2	
Trp	-	-	
Tyr	1	-	
His-Asn	-	1	
Glu-Lys	1	-	
Glu-Asp	1	-	
Met-Asn	1	-	6
Phe-Leu	1	-	
Ser-Glu	1	-	
Asn-Asp	1	-	
Ser-Asn	1	-	
Diverse	1	-	4
Total	**159**	**74**	

[#] includes 5 hexanucleotides
[+] includes 1 mononucleotide, [++] includes 2 mononucleotides

Almost all homopolymers were encoded by trinucleotides and a small number of mono- and hexanucleotides. The heteropolymers were encoded by tetra- and hexanucleotides. Above all glutamic acid and glutamine were the most abundant homopolymers followed by asparagine, aspartic acid and serine repeats. Interestingly, amongst representative homopolymers (>9 loci) alanine, asparagine, glutamine, and phenylalanine are frequently polymorphic whereas stretches of, for example, aspartic acid appear only mildly polymorphic.

Next, we examined the functions of these genes and genes corresponding to promoter regions containing microsatellites using GO term annotations (see methods). At a whole genome scale, genes containing microsatellites are mostly associated with the regulation of biological and cellular processes, especially transcription (regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process) (Table 4). On the other hand, the subset of conserved microsatellites are also associated with genes involved in the regulation of biological and cellular processes, but are not involved in transcription despite a remaining association with RNA metabolic processes. Further they show a small association with growth, actin cytoskeleton organization, sexual reproduction, and conjugation. More specifically, genes containing variable conserved microsatellites exhibit unique associations with cellular structure morphogenesis and responses to pheromones. In contrast, monomorphic (non-variable) conserved microsatellites appear in genes with the exact same functions as conserved and genomic microsatellites. The differences between monomorphic and polymorphic microsatellite indicate different selective pressures acting on the associated genes.

The search yielded no significant functional clusters for genes corresponding to promoter regions containing either genomic microsatellites or polymorphic conserved microsatellites because a large number of genes were annotated with unknown function.

Table 4. Significant functional annotations (GO term clusters) for genes containing microsatellites [(p>0.01), background: entire gene set of *S. cerevisiae*]

GO term (Process)	Back-ground (%)	Fraction (%) of genes with microsatellites			
		all perfect (genomic)	all conserved	mono-morphic	poly-morphic
actin cytoskeleton organization and biogenesis	1.5		6		
biological regulation	13.1	**23.9**	24.9		
cell communication	3.5	6.5			
cell cycle	6.2	10.8			
cell cycle phase	4.8	8.4			
cell cycle process	5.5	9.5			
cell growth	1.2	3.1	5.6		
cellular component organization and biogenesis	30.5	**41.9**	45.5		55.8
cellular process	64.1	71.3			
cellular structure morphogenesis, anatomical structure development	3.5				15.6
chromatin modification	3	5.9			
chromosome organization and biogenesis	7.9	12.9			
conjugation with cellular fusion	1.6		6		
establishment and/or maintenance of chromatin architecture	3.5	6.6			
filamentous growth	1.3	3.3			
growth	2	5.1	**8.2**	9.2	
mitotic cell cycle	3.7	**7**			
multi-organism process			6.9		
negative regulation of biological process	3.6	6.6			
negative regulation of cellular process	3.5	6.6			
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	23.5	**31.6**			
organelle organization and biogenesis	19.8	**28.9**	**34.3**		
regulation of biological process	10.7	**20.2**	**23.6**	23.7	
regulation of biological quality	4	7.1			
regulation of cell cycle	2.3	**5.5**			
regulation of cell size	1.6	**4.1**			
regulation of cellular metabolic process	7.4	13.7			
regulation of cellular process	10.5	**19.7**	**23.6**	23.1	
regulation of gene expression	6.1	**11.3**			
regulation of metabolic	7.6	**14.3**			

process				
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6.3	**11.9**		
regulation of RNA metabolic process	5.4	**10.4**		
regulation of transcription	5.5	**10.3**		
regulation of transcription from RNA polymerase II promoter	3.2	6		
regulation of transcription, DNA-dependent	5.1	**9.4**		
response to abiotic stimulus	1.6	3.8		
response to pheromone	1.3			10.4
response to stimulus	10.9	**16.8**		
RNA biosynthetic process	7.3	13.2		
RNA metabolic process	14.6	**21.1**	24.9	
sexual reproduction	1.6		6	
transcription	7.9	**15.2**		
transcription from RNA polymerase II promoter	4.9	**9.4**		
transcription, DNA-dependent	7.3	**13.2**		

* p< 0.0001 in bold and underlined

We then asked whether this distribution could be an artefact of the overall conservation of genes, and if microsatellites just had "come along for the ride", i.e. only persist as part of a conserved gene. Overall genes containing conserved microsatellite show on average higher synonymous substitution rates than the genomic background (Table 5, Figure 7). Genes containing variable microsatellites exhibit low rates of non-synonymous substitutions.

Table 5. Mean evolutionary rates for various genes sets (from (Hirsh *et al.* 2005))

Gene set	ORF	dS'	dN	dN/dS'
All genes	3036	2.136	0.167	0.077
Genes with conserved SSR	112	**2.205 *(p=0.002)***	0.175 (*p=0.476*)	0.079 (*p=0.734*)
Genes with monomorphic SSR	81	**2.233 (*p<0.001*)**	0.173 (*p=0.611*)	0.077 (*p=0.986)*
Genes with polymorphic SSR	38	2.126 (*p=0.805*)	**0.164 (*p<0.001*)**	0.077 (*p=0.984*)

*approximate p-value estimated via non-parametric permutation test

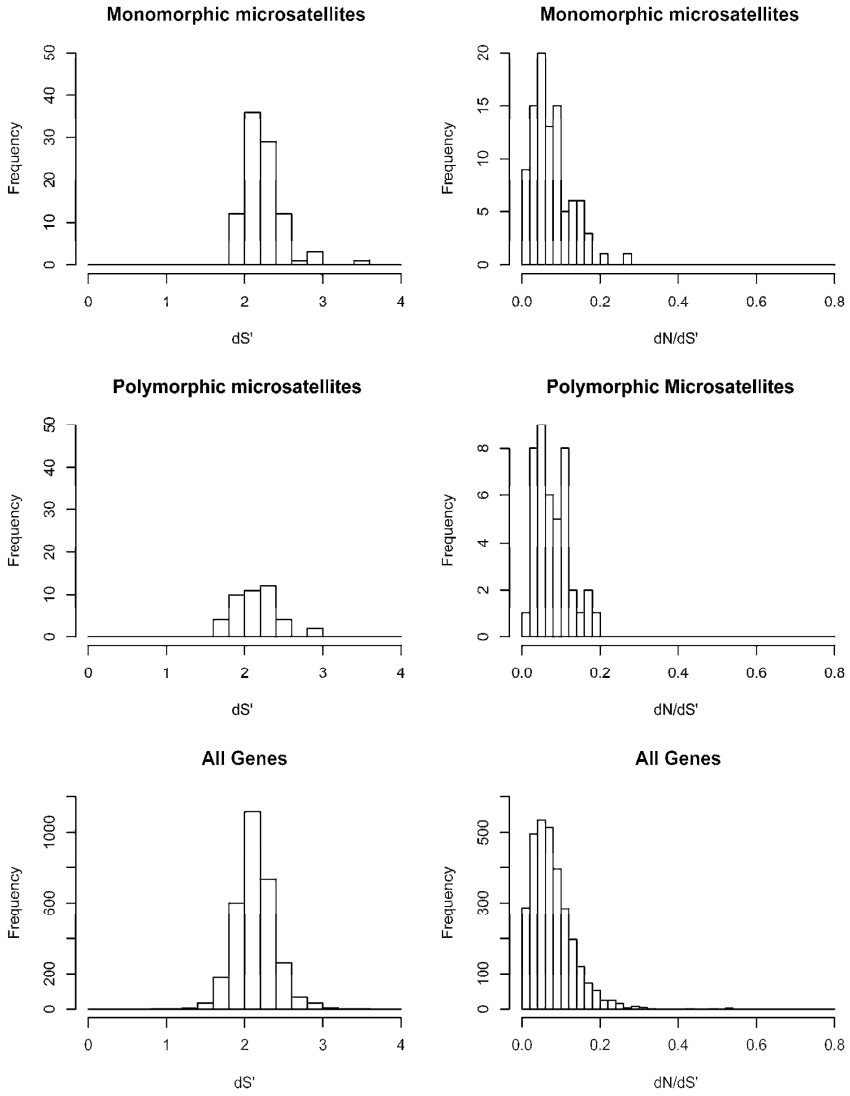


Figure 7. Evolutionary rates for genes with conserved microsatellites (monomorphic, polymorphic) and the genomic set as estimated by Hirsh et al (Hirsh *et al.* 2005). dS' = synonymous nucleotide substitutions per synonymous site (adjusted for codon adaptation index); dN = non-synonymous substitution per non-synonymous site

5.4. Discussion

Microsatellites persist predominantly in regions of the genome where their polymorphism is not deleterious and where overall substitutions (i.e. SNP density) are low, allowing for

replication slippage and the conservation of flanking sequences. As such microsatellites are conserved as a result of their genomic position: non-coding regions allow for high microsatellite frequencies (starting from mononucleotide and decreasing towards hexanucleotide repeats), whereas coding regions restrict microsatellite abundance to a few non-deleterious trinucleotides. The negative influence of coding sequence (CDS) on microsatellite frequencies is evident in the distribution of microsatellites in most genomes (Toth *et al.* 2000; Katti *et al.* 2001; Dieringer and Schlotterer 2003) and favours neutral behaviour for the majority of loci. Interestingly, the loss of microsatellite loci in promoter regions immediately adjacent to CDS is similar in magnitude to the loss in CDS itself, which indicates linkage and demonstrates a wider effect of CDS on microsatellite distribution than the ORF limits *per se*. Further, the observed negative correlation between SNP density and microsatellite frequency supports the balanced replication-slippage model for microsatellite evolution proposed by Kruglyak et al. (1998(Kruglyak *et al.* 1998; Kruglyak *et al.* 2000) (see *Chapter 4*). Under this model, microsatellite mutations are governed by a length dependent replication slippage on one hand, and point mutations, i.e. imperfections in the array, which lower slippage rate, on the other. Concurrently, a similar observation has been recently made in the chicken genome (Brandstrom and Ellegren 2008).

Proportions of polymorphic loci across coding and non-coding regions are very similar (at least for mono- to trinucleotides which constitute the majority, i.e. ~ 95%, of microsatellite loci), which excludes a global effect of transcription on microsatellite polymorphism that had been suggested before (Wierdl *et al.* 1996). However, such similar occurrences in polymorphism are still somewhat surprising, since evolutionary forces, such as the selection imposed on loci due to functional constraints, act quite differently on coding and non-coding regions (Page and Holmes 1998). Additionally, in contrast to other studies, the proportions of polymorphic loci as well as distributions of polymorphism (K/N) are very similar across different repeat unit sizes. Previous studies based on population data and genomic length distributions in humans (Chakraborty *et al.* 1997), fruit fly (Schug *et al.* 1998) and yeast (Kruglyak *et al.* 2000) have estimated

mutation rates that were (on average) inversely related to the repeat unit size, i.e. highest for di-, followed by tri-and tetranucleotides

It is likely, however, that the evolutionary divergence amongst strains is not sufficient to create the amount of microsatellite polymorphism needed to enable differentiation (268 polymorphic loci in total, 86 loci ≥ 3 alleles, 22 loci ≥ 4 alleles). We also excluded several polymorphic loci during the filtering process when selecting for unique loci (for details see methods). For example, earlier attempts to identify polymorphic microsatellites for the purpose of yeast strain identification had proposed a total of 41 loci, of which the 6 most polymorphic could be amplified across 47 strains (Legras *et al.* 2005). Comparing those 6 loci with our genomic and conserved sets yielded full overlap with the genomic set, but only one out of six loci (YPL009c) was detected within the conserved set - all five loci were lost as part of a larger chromosomal segment. Nevertheless, we still detect the characteristic positive relationship between microsatellite polymorphism and array length, which confirms that the predicted polymorphism of markers is non-random and reflects the dominating influence array length has on microsatellite variability compared to other factors (Wierdl, Dominska et al. 1997; Brinkmann, Klintschar et al. 1998; Primmer, Saino et al. 1998; Schug, Hutter et al. 1998).

Owing to its compactness (~70% coding sequence), the yeast genome comprises a high frequency of trinucleotides repeats, much higher than those found in mammalian genomes (Toth *et al.* 2000). Although, we find the majority of motifs experience frequency dependent conservation, a few motifs experience a turn-over which is slower or faster than the average. Previous studies have already noted a codon bias for triplet repeats in yeast where repeats of glutamine, asparagine, aspartic acid, glutamic acid and serine are present in higher numbers than expected, preferentially in regulatory genes and transcription factors (Alba *et al.* 1999; Young *et al.* 2000). This trend extends across most eukaryotic species (Alba *et al.* 1999; Richard *et al.* 1999; Karlin *et al.* 2002; Alba and Guigo 2004; Faux *et al.* 2005) and has been explained by the underlying strand slippage mutation mechanism and the propensity of these repeats to form secondary structures further enhancing the likelihood of slippage events (Gacy *et al.* 1995).

Our results from 40 *S.cerevisiae* strains show that these amino acids repeats form persisting loci indicating functionality. Concordant patterns have been observed in 12 *Drosophila* species (Huntley and Clark 2007), human-mouse orthologs (Mularoni *et al.* 2007) and rat-mouse-human orthologs (Alba and Guigo 2004). Not much is known about the functional implications of homopolymers, but it has been suggested that the increased flexibility of unstructured domains created by homopolymers may mediate the assembly of large protein complexes (Faux *et al.* 2005). On the other hand, polyglutamine and polyalanine stretches feature several human genetic diseases that are caused by the insoluble protein aggregates, modulated protein-protein interactions and/or altered gene expression (Brown and Brown 2004; Gatchel and Zoghbi 2005).

We find that genes containing conserved microsatellites show elevated mutation rates compared to the genomic background, which suggests that microsatellites are more easily tolerated in a more variable environment. A similar observation has been recently reported for homopolymers in Drosophila and human-mouse orthologs (Huntley and Clark 2007; Mularoni *et al.* 2007). However, we also find that polymorphic loci are located within genes that show reduced rates of non-synonymous substitutions. On one hand, this could be due to codon redundancy or, on the other, may indicate the influence of selective forces. The phenotypic effects created by variable microsatellites in coding and regulatory regions, particularly those believed to provide an evolutionary advantage for the organism, have led to the anaphor of 'evolutionary tuning knobs' for microsatellites (Kashi and King 2006). We find that beside a prevailing association with cellular component organization and biogenesis a few polymorphic loci are associated with cell morphogenesis, anatomical structure development and pheromone response during mating, budding or as a response to stress. Mutations in those genes could alter reproductive efficiency and cell growth, which may modulate global metabolic and physiological activity and, considering the variety of habitats for *S. cerevisiae*, could provide an adaptive advantage under varying environmental conditions. For example, two of these genes, *CDC39* (basal transcription factor, 3'–5'-exoribonuclease) and *NPL3* (m-RNA binding protein), are directly associated with pseudohyphal growth – a virulence trait that *S. cerevisae* exhibits as an opportunistic human pathogen (McCusker *et al.*

1994; Hazen 1995). However, we find no evidence for the previously proposed hypothesis that intragenic microsatellite variability might be associated with variation in the expression of cell surface molecules. *FIG2* is the only the cell wall adhesion we identified amongst the latter group, but is expressed specifically during mating and can not account for an antigen diversity. Finally, we do have to note, that despite no significant association of intragenic polymorphic microsatellites with transcription, we still find several RNA-binding proteins containing this type of microsatellites.

Despite these very plausible indications, actual phenotypic effects can only be validated experimentally and further work in this area is needed. Of particular interest, we found that several polymorphic loci where conserved across the *sensu strictu* (*Saccharomyces*) group (for preliminary data see Appendix). It would be interesting to extend this study across a larger evolutionary time scale, such as the entire *Hemiascomycetes* phylum, to further identify candidate functional microsatellite loci.

5.5. Summary

Microsatellites appear as neutrally evolving sequences, i.e. that they are conserved as a result of their genomic position. Their evolution dynamics follow a counter play between polymorphism creating strand slippage on one hand and stabilizing point mutations on the other. Repeat array length is positively correlated with polymorphism. Some microsatellites, however, experience a slower turnover than the average. In coding regions, the conservation of microsatellite sequences could have several explanations: 1) a function of unstructured protein domains created by homopolymers in protein-protein or protein-nucleic-acid interactions, 2) a tolerance for microsatellite repeats in genes under low selective constraints, or 3) functional importance of variable microsatellites in establishing phenotypic diversity resulting in an evolutionary advantage.

Chapter 6

Summary, Discussion and Future Directions

Microsatellites are a common element of most genomes and have been used extensively as genetic markers over the last two decades (Schlotterer 2004) . Their mutational dynamics have been proven to be complex and still retain many areas of contention (Ellegren 2004). Microsatellites are non-randomly distributed throughout the genome, associated with other sequence features and located within genes (or thereby regulatory regions) comprising distinct functional groups (Li *et al.* 2002; Morgante *et al.* 2002; Malpertuy *et al.* 2003; Alba and Guigo 2004). Human genetic disorders caused by excessive trinucleotide expansions and adaptive evolution in pathogens through contingency loci are the most well known phenotypic effects of microsatellite polymorphism, but the number of studies reporting effects of microsatellites are increasing (Li *et al.* 2004; Gatchel and Zoghbi 2005; Kashi and King 2006). Could those simple sequences possibly possess a biological meaning beyond simple neutral genetic markers?

6.1. The genomic age – new data, new methods, old pitfalls ?!

The genomic age has brought about an ever-increasing amount of genomic data that may hold the clue to deciphering the mysteries of microsatellite evolution. But it has also raised old discrepancies of studying microsatellites within the newly developed *in silico* methods. For example, ever since their discovery 20 years ago, researchers have selected and defined microsatellite sequences through their research focus, rather than the biological characteristics of the sequences themselves. So microsatellites are generally described as a short tandemly repeated DNA motif, but the minimum number of repeats, as well as, the actual length of the repeated motif, or the type of repetition (perfect or imperfect) is often set arbitrarily by the researcher. However, these characteristics

themselves are in fact major influences of microsatellite mutation and genomic distribution (see General Introduction: "Factors influencing microsatellite mutation"), e.g. microsatellite abundance increases exponentially with decreasing array length, though mutation rates are higher for longer arrays. Our meta-study in *Chapter 2* shows that because of the arbitrary definitions set, estimates of genomic microsatellite distribution even in a single genome significantly diverge. At the extreme, we find a divergence in the order of several magnitudes between two studies investigating microsatellite distribution in yeast, but also observe differences in the relative frequencies of individual size classes (e.g. higher/ lower frequency of mononucleotides compared to trinucleotides; see Figure 1, pg. 22). While such discrepancy is a concern to the immediate conclusions drawn from an individual study, the disjunct in the results reported by different studies on the same genome is particularly worrying and draws into question the validity of cross-species comparisons undertaken using published data, i.e. study comparisons might not actually be variable due to different selected (sub-) types of microsatellites, and species-specific patterns might not be attributable to species-specific factors. Consensus in approach is strongly needed, particularly for genomic studies due to the sheer number of loci involved.

Another factor that might compromise genomic comparisons is variance in the choice of analysis tool. Throughout the last decade genomic studies have produced a vast range of bioinformatic tools to detect and analyze microsatellites in genomic data (e.g. Abajian 1994; Benson 1999; Kolpakov *et al.* 2003). Until recently it appeared that the scientific community had not caught up with such diversity, since the description of the employed tools were largely limited or unavailable. To motivate a selective tool choice over applying software as a black box, I reviewed the existing range of microsatellite detecting tools (*Chapter 3*) describing the underlying algorithms, their search strategy, efficiency, utility and suitability for certain research purposes. The tools differed remarkably in all of these characteristics and selecting a tool is best achieved based on the users prior experience, computational resources and the purpose of the analysis, which will impact on the choices for downstream data analyses. Within *in silico* studies, inconsistent microsatellite definitions directly manifest as multiple parameter settings input by the

user or are embedded as individual search approaches in the algorithm. Both create again study biases and lead to large discrepancies in outcomes (see also Leclercq *et al.* 2007). Significant results should therefore ideally be evaluated using a second approach and more importantly, details of the approach, i.e. a formal description of the search tool and parameter settings (see Box 6.1., *Chapter 2, 3*), need to be provided in order to allow a sensible comparison amongst studies. Some studies have adapted these suggestions (e.g. Kelkar *et al.* 2008), but the importance of such information cannot be stressed enough since the flow of genomic data is only about to increase and with it the number of microsatellites studies.

Box 6.1.: Critical information need for an *in silico* study of STRs (short tandem repeats)

1. Type of repeat: perfect, imperfect, compound/ complex
2. Software (ideally a short description of the search algorithm)
3. Search parameters: minimum array length; motif length; number of matches, mismatches and/ or indels allowed
4. Filter: duplicates, overlapping repeats, motifs

6.2. Microsatellites as genomic entities

Many studies of microsatellite evolution have focused on the contribution of cellular processes like replication, recombination and repair to microsatellite variability, i.e. the underlying mechanism(s) of microsatellite mutation. Most investigations were related to intrinsic factors influencing microsatellite mutation rate such as motif length and type, internal structure, and particularly array length (see *General Introduction* for examples). A number of studies also found a preference for microsatellites to be associated with particular transposable elements (*Alu, LINEs* and *SINEs* in humans, *mini-me* in

Dipterans) and suggested a connection with their origin (Nadir *et al.* 1996; Wilder and Hollocher 2001). But besides the observed distinction between coding and non-coding microsatellites, there have been comparatively few studies examining relationships between microsatellites and their immediate genomic environment, especially on a genome-wide level. The increasing amounts of genomic data, particularly functional and structural data, is about to reverse that trend. For example, in a very recent study, Kelkar et al. (2008) used orthologous loci from genomic human-chimp alignments, to compare the influences of locus related factors vs. genomic factors (e.g. locations in different isochores, local (>1Mb) GC-content, recombination frequency) on microsatellite mutability, and found, despite significance in both groups, a prevailing dominance of the former. Others have found weak associations with recently mapped recombinational hotspots in yeast and humans (Bagshaw 2008)

We find that microsatellites are indeed sensitive to their genomic neighborhood or, conversely, a reflection of it. In our study in yeast (*Chapter 4*), we see microsatellite depletion in coding regions, regions nearby LTR transposons and regions of high SNP density. Amongst the latter instances, depletion of microsatellites likely mark recently expanded genome regions on one hand (Morgante *et al.* 2002) but display the negative stabilizing influence of substitutions (imperfect repeat copies) on their polymorphism on the other (Petes *et al.* 1997; Rolfsmeier and Lahue 2000), respectively. In some cases the association with proximate sequence features indicates functional implications. Poly(A) or $(AAT)_n$ are associated with meiotic double-strand breaks and nucleosomes, respectively; whereas regulatory sites show spatial correlation with poly(A) and $(AT)_n$ motifs. This pattern could be the result of some property of the microsatellite sequence, i.e. some alteration of DNA topology, or a functional implication connected to their variability (Schultes and Szostak 1991; Gendrel *et al.* 2000; Suter *et al.* 2000; Kashi and King 2006). In other cases, as GC-content analysis in our study shows, motif frequency is in coherence with the local background composition, which supports a neutral model of evolution for these microsatellites.

In any case, our results and others (Bagschaw 2008; Bagshaw 2008; Kelkar *et al.* 2008) show that microsatellites do act as genomic entities and are susceptible to local influences. Locus related factors (array length, motif type and length, degradation of the array) and negative selection might be the prevailing factors influencing overall microsatellite mutation rate and genomic distribution, but different factors act on different scales. Although local influences are subtle in the larger picture of microsatellite evolution, they will become rapidly significant when we start considering different populations of microsatellites within the genome and, of course, when examining individual loci.

6.3. Microsatellite functions

Microsatellite sequences can make up a substantial proportion of the genome. In yeast we estimated a genome coverage of 0.3%, but in repeat-rich mammalian genomes coverage can be up to 5 % (Warren *et al.* 2008) which, for example, in humans (3% genome coverage) exceeds the genome coverage of protein coding genes (1.5%) (Lander *et al.* 2001). The striking number of microsatellites within the genome has made it likely that amongst the bulk of neutrally evolving loci, a few loci might acquire a function. For example, in *Chapter 3* we have shown that this could involve structural sequence features such as meiotic double strand breaks or nucleosomes. Secondly, microsatellites have been considered as 'evolutionary tuning knobs' (Kashi *et al.* 1997) due to their ability to regulate gene expression through frequent and reversible mutations resulting in advantageous phenotypic diversity. Others have suggested that homopolymers in proteins (i.e. intragenic microsatellites) might fulfill a function as flexible protein domains in modulating protein-protein or protein-nucleic-acid interaction (Faux *et al.* 2005).

In comparative genomics sequence conservation is generally an indicator for functionality. However, it is difficult to differentiate between selection and neutral mutation. I have taken an exploratory approach analyzing the conservation and polymorphism of microsatellites across 40 sequenced yeast strains to identify signals for functionality on a genome wide scale (*Chapter 5*). Unsurprisingly, I found overwhelming

evidence of neutral mutational behavior for microsatellites, such as the predominant persistence and variability of microsatellites in regions of low evolutionary constraints (i.e. non-coding regions and genes that show elevated mutation rates). Further, individual motifs were conserved in a frequency dependent manner, which also supports neutral evolution. Nevertheless, since not all motifs occur at similar frequencies throughout the genome, this coincides with a previously reported amino acid bias in coding regions (Alba *et al.* 1999 ; Young *et al.* 2000; Alba and Guigo 2004). On one hand this bias could indicate some function of homopolymers (Faux *et al.* 2005), or alternatively, be explained by the heightened propensity of individual motifs to form secondary structures and undergo increased numbers of mutational events (Wells *et al.* 2005). We find, genes containing persistent microsatellites and particularly polymorphic microsatellites are associated with the organization and biogenesis of cellular components, morphogenesis, development of anatomical structures and pheromone response (see *Chapter 5*, Table 4). This again, points towards a source of phenotypic variation with some adaptive advantage (Kashi and King 2006), or conversely, connects back to our observation that those genes experience low evolutionary constraints where it is feasible to believe that microsatellite mutations are more easily tolerated. However, a very recent study on conserved amino acid repeats in *Drosophila* has found evidence that repeat containing genes show accelerated evolution, suggesting influences of positive selection (Huntley and Clark 2007).

Contributing to the debate over neutralityversus selection as the predominant force acting on genome evolution and particularly pervasive repetitive sequences such as microsatellites is work published by Michael Lynch and others (e.g. Lynch and Walsh, 1998; Lynch *et al* 2003) that suggests that effective population size has a significant impact on genome architecture. Lynch argues that small population size (Ne) reduces the efficiency of selection and permits the acquisition of slightly deleterious mutations such as the expansion of repetitive sequences. Among his key evidence is a strong negative relationship between Ne, genome size and the number of transposable elements (Lynch *et al* 2003). As a result of this process sequence material is created that can be coopted into new biological functions. Through their characteristics, such as ubiquitous abundance and

high mutation rates, microsatellite sequences are the perfect contestants for such theories. In fact, microsatellite frequencies in species with small effective population size such as mammals tend to be much higher than frequencies observed in unicellular eukaryotes with large population sizes like yeasts. A promising future study to efficiently test for such a correlation could be easily based around Lynch's paper on the "Origins of Genome Complexity" (Lynch *et al* 2003) where he derives population size estimates for a range of species from measures of silent-site nucleotide variation. In a similar fashion one could investigate turn-over rates (gain and loss) of microsatellite loci within different taxa.

6.4. The future of microsatellites

For over 20 years, microsatellites have been used frequently as genetic markers for a variety of applications. Recently, however, their popularity seems in decline (Schlotterer 2004). Their heterogeneous mutation rates have created difficulties for population genetic studies, since the currently utilized mutation models do not capture their complex mutational dynamics sufficiently which is particularly important when transferring genetic distances into absolute time scales (Balloux and Lugon-Moulin 2002; Landry *et al.* 2002). The development of high-throughput analyses and finer density genomic data for an increasing number of species has further advanced the use of SNPs as molecular markers for association studies. Despite the larger information content of a single microsatellite locus, genomic SNP and maps outweigh those gained from microsatellites (Dunn *et al.* 2005; Hinrichs *et al.* 2005). Nevertheless, the use of multiple loci can overcome the heterogeneity amongst loci and, compared with the shot-gun sequencing approach necessary for SNPs, microsatellites are still the cheaper option for marker development in non-model organisms (Schlotterer 2004). Further, considering our and other recent results (Brandstrom and Ellegren 2008), microsatellites could still outweigh the utility of SNPs in regions of low SNP density due to the inverse correlation between microsatellite variability and SNP density.

However, there are still opportunities for new applications for microsatellites. Recently microsatellites have been shown to be particularly useful in hitchhiking mapping studies, whereby genome wide microsatellite variability is utilized to identify genomic regions that have been targeted by selection (Schlotterer 2003). Alternatively, microsatellite depletion could be used to identify recent genome expansions, since their high mutation rate would allow them to quickly 'colonize' newly founded genomic regions (e.g. as a result of LTR transposon activity, see *Chapter 3,* also (Morgante *et al.* 2002)).

Despite our (and others) very plausible indications for variable microsatellites altering gene expression and facilitate adaptive evolution, actual phenotypic effects can only be validated experimentally. QTL and association studies identify 'neutral' variants (e.g. microsatellite alleles) that are linked to a certain phenotype on a large scale, but, due to their nature, but cannot provide any support for an underlying mechanistic connection. *In vitro* and *in vivo* approaches employing reporter-gene assay systems, on the other hand, can directly measure (and compare) gene expression for different alleles, but are limited to a small number of loci. Nevertheless work in this area is strongly needed.

Alternatively, an extension of the comparative genomics approach across larger evolutionary divergence, i.e. the *sensus strictu* (*Saccharomyces*) group or the *Hemiascomycetes* phylum would provide stronger evidence for influences of selection. So called 'phylogenetic footprinting', i.e. identifying functional elements by means of sequence conservation, has been successfully used to identify functional elements in a variety of taxa including yeasts (Cliften *et al.* 2003) and also mammals (Dermitzakis and Clark, 2002). It also has shown that regulatory sites, such as transcription factor binding sites, can undergo frequent turn-over (Dermitzakis and Clark, 2002).

References

Aaltonen, L. A., P. Peltomaki, et al. (1993). Clues to the Pathogenesis of Familial Colorectal-Cancer. *Science* **260**(5109): 812-816.

Abajian, C. (1994). Sputnik. http://espressosoftware.com/pages/sputnik.jsp

Aishwarya, V., A. Grover, et al. (2007). EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics* **8**: 225.

Aho, A. V. and M. J. Corasick (1975). Efficient string matching: and aid to bibliographic search. *Communications of the ACM* **18**(6): 333-340.

Ak, P. and C. J. Benham (2005). Susceptibility to Superhelically Driven DNA Duplex Destabilization: A Highly Conserved Property of Yeast Replication Origins. *PLoS Computational Biology* **1**(1): 41-46.

Alba, M. M. and R. Guigo (2004). Comparative analysis of amino acid repeats in rodents and humans. *Genome Research* **14**(4): 549-554.

Alba, M. M., R. A. Laskowski, et al. (2002). Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* **18**(5): 672-8.

Alba, M. M., M. F. Santibanez-Koref, et al. (1999). Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. *Journal of Molecular Evolution* **49**(6): 789-797.

Amos, W., C. M. Hutter, et al. (2003). Directional evolution of size coupled with ascertainment bias for variation in Drosophila microsatellites. *Molecular Biology and Evolution* **20**(4): 660-662.

Amos, W., S. J. Sawcer, et al. (1996). Microsatellites show mutational bias and heterozygote instability. *Nature Genetics* **13**(4): 390-391.

Anderson, J. D. and J. Widom (2001). Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Molecular and Celluar Biology* **21**(11): 3830-9.

Arcot, S. S., Z. Wang, et al. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics* **29**(1): 136-44.

Areshchenkova, T. and M. W. Ganal (1999). Long tomato microsatellites are predominantly associated with centromeric regions. *Genome* **42**(3): 536-44.

Armour, J. A. L., S. A. Alegre, et al. (1999). Minisatellites and mutation processes in tandemly repetitive DNA. *Microsatellites: Evolution and Applications*. D. Goldstein and C. Schlötterer. New York, Oxford University Press**:** 24-33.

Bachtrog, D., S. Weiss, et al. (1999). Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. *Molecular Biology and Evolution* **16**(5): 602-610.

Bagshaw, A. T. (2008). An investigation of links between simple sequences and meiotic hotspots. *School of Biological Sciences*. Christchurch, University of Canterbury. **PhD**.

Bagshaw, A. T., J. P. Pitt, et al. (2008). High frequency of microsatellites in S. cerevisiae meiotic recombination hotspots. *BMC Genomics* **9**(1): 49.

Baldi, P. and P. F. Baisnee (2000). Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* **16**(10): 865-889.

Balloux, F. and N. Lugon-Moulin (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**(2): 155-165.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(2): 573-580.

Birdsell, J. A. (2002). Integrating Genomics, Bioinformatics, and Classical Genetics to Study the Effects of Recombination on Genome Evolution. *Molecular Biology and Evolution* **19**(7): 1181-1197.

Bizzaro, J. W. and K. A. Marx (2003). Poly: a quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. *Bmc Bioinformatics* **4**: 22.

Boeva, V., M. Regnier, et al. (2006). Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* **22**(6): 676-684.

Borstnik, B. and D. Pumpernik (2002). Tandem repeats in protein coding regions of primate genes. *Genome Research* **12**(6): 909-915.

Botstein, D., S. A. Chervitz, et al. (1997). Yeast as a model organism. *Science* **277**(5330): 1259-1260.

Boyle, E. I., S. Weng, et al. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**(18): 3710-5.

Brandes, A., H. Thompson, et al. (1997). Multiple repetitive DNA sequences in the paracentromeric regions of Arabidopsis thaliana L. *Chromosome Res* **5**(4): 238-46.

Brandstrom, M. and H. Ellegren (2008). Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Research* **18**(6): 881-887.

Brinkmann, B., M. Klintschar, et al. (1998). Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* **62**(6): 1408-1415.

Brock, G. J. R., N. H. Anderson, et al. (1999). Cis-acting modifiers of expanded CAG CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. *Human Molecular Genetics* **8**(6): 1061-1067.

Brohede, J. and H. Ellegren (1999). Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proceedings of the Royal Society of London Series B-Biological Sciences* **266**(1421): 825-833.

Brown, L. Y. and S. A. Brown (2004). Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends in Genetics* **20**(1): 51-8.

Brown, T. A. (1999). *Genomes*. Oxford, BIOS Sciebtific Publishers Ltd.

Brown, T. C. and J. Jiricny (1989). Repair of base-base mismatches in simian and human cells. *Genome* **31**(2): 578-83.

Buard, J. and A. J. Jeffreys (1997). Big, bad minisatellites. *Nature Genetics* **15**(4): 327-8.

Buhler, C., V. Borde, et al. (2007). Mapping Meiotic Single-Strand DNA Reveals a New Landscape of DNA Double-Strand Breaks in *Saccharomyces cerevisiae*. *PLoS Biology* **5**(12): e324.

Bull, L. N., C. R. Pabon-Pena, et al. (1999). Compound microsatellite repeats: Practical and theoretical features. *Genome Research* **9**(9): 830-838.

Buschiazzo, E. and N. J. Gemmell (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays* **28**(10): 1040-1050.

Cao, H., H. R. Widlund, et al. (1998). TGGA repeats impair nucleosome formation. *Journal of Molecular Biology* **281**(2): 253-260.

Caporale, L. H. (2003). Natural Selection and the Emergence of a Mutator Phenotype: An Update of the Evolutionary Synthesis Considering Mechanisms that Affect Genome Variation. *Annual Review of Microbiology* **57**(1): 467-485.

Castelo, A. T., W. Martins, et al. (2002). TROLL-Tandem Repeat Occurrence Locator. *Bioinformatics* **18**(4): 634-636.

Chakraborty, R., M. Kimmel, et al. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences of the United States of America* **94**(3): 1041-1046.

Chambers, G. K. and E. S. MacAvoy (2000). Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **126**(4): 455-476.

Charlesworth, B., M. T. Morgan, et al. (1993). The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics* **134**(4): 1289-1303.

Chen, Y. and R. Roxby (1997). Identification of a functional CT-element in the Phytophthora infestans piypt1 gene promoter. *Gene* **198**(1-2): 159-164.

Cliften, P., P. Sudarsanam, et al. (2003). Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**(5629): 71-76.

Colson, I. and D. B. Goldstein (1999). Evidence for complex mutations at microsatellite loci in Drosophila. *Genetics* **152**(2): 617-627.

Coote, T. and M. W. Bruford (1996). Human microsatellites applicable for analysis of genetic variation in apes and old world monkeys. *Journal of Heredity* **87**(5): 406-410.

Delgrange, O. and E. Rivals (2004). STAR: An algorithm to search for tandem approximate repeats. *Bioinformatics* **20**(16): 2812-2820.

Depledge, D. P. and A. R. Dalby (2005). COPASAAR--a database for proteomic analysis of single amino acid repeats. *BMC Bioinformatics* **6**: 196.

Dermitzakis, E. T. and A. G. Clark (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Molecular Biology and Evolution* **19**(7): 1114-1121.

Dieringer, D. and C. Schlotterer (2003). Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. *Genome Research* **13**(10): 2242-2251.

Dirienzo, A., A. C. Peterson, et al. (1994). Mutational Processes of Simple-Sequence Repeat Loci in Human-Populations. *Proceedings of the National Academy of Sciences of the United States of America* **91**(8): 3166-3170.

Duffy, A. J., D. W. Coltman, et al. (1996). Microsatellites at a common site in the second ORF of L1 elements in mammalian genomes. *Mammalian Genome* **7**(5): 386-7.

Dujon, B. (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends in Genetics* **22**(7): 375-387.

Dunn, G., A. L. Hinrichs, et al. (2005). Microsatellites versus single-nucleotide polymorphisms in linkage analysis for quantitative and qualitative measures. *BMC Genetics* **6**(Suppl.1).

Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics* **24**(4): 400-402.

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* **16**(12): 551-558.

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics* **5**(6): 435-445.

Ellegren, H., G. Lindgren, et al. (1997). Fitness loss and germline mutations in barn swallows breeding in Chernobyl. *Nature* **389**(6651): 593-596.

Faux, N. G., S. P. Bottomley, et al. (2005). Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Research* **15**(4): 537-551.

Field, D. and C. Wills (1998). Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and S-cerevisiae, result from strong mutation pressures and a variety of selective forces. *Proceedings of the National Academy of Sciences of the United States of America* **95**(4): 1647-1652.

Fitzpatrick, D. A., M. E. Logue, et al. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology* **6**: 99.

FitzSimmons, N. N., C. Moritz, et al. (1995). Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Molecular Biology and Evolution* **12**(3): 432-40.

Fondon, J. W. and H. R. Garner (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America* **101**(52): 18058-18063.

Freudenreich, C. H., J. B. Stavenhagen, et al. (1997). Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Molecular and Cellular Biology* **17**(4): 2090-2098.

Gacy, A. M., G. Goellner, et al. (1995). Trinucleotide Repeats That Expand in Human-Disease Form Hairpin Structures in-Vitro. *Cell* **81**(4): 533-540.

Gatchel, J. R. and H. Y. Zoghbi (2005). Diseases of unstable repeat expansion: Mechanisms and common principles. *Nature Reviews Genetics* **6**(10): 743-755.

Gendrel, C. G., A. Boulet, et al. (2000). (CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis. *Genes & Development* **14**(10): 1261-1268.

Gerber, H. P., K. Seipel, et al. (1994). Transcriptional Activation Modulated by Homopolymeric Glutamine and Proline Stretches. *Science* **263**(5148): 808-811.

Gerton, J. L., J. DeRisi, et al. (2000). Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America* **97**(21): 11383-90.

Giardine, B., C. Riemer, et al. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* **15**(10): 1451-1455.

Glenn, T. C., W. Stephan, et al. (1996). Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. *Molecular Biology and Evolution* **13**(8): 1151-1154.

Goffeau, A., B. G. Barrell, et al. (1996). Life with 6000 genes. *Science* **274**(5287): 546-567.

Goldstein, D. B. and A. G. Clark (1995). Microsatellite Variation in North-American Populations of Drosophila-Melanogaster. *Nucleic Acids Research* **23**(19): 3882-3886.

Goldstein, D. B. and D. D. Pollock (1997). Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. *Journal of Heredity* **88**(5): 335-342.

Goldstein, D. B. and C. Schlötterer (1999). *Microsatellites : Evolution and Applications*. Oxford ; New York, Oxford University Press.

Gomes-Pereira, M., M. T. Fortune, et al. (2001). Mouse tissue culture models of unstable triplet repeats: in vitro selection for larger alleles, mutational expansion bias and tissue specificity, but no association with cell division rates. *Human Molecular Genetics* **10**(8): 845-854.

Hammock, E. A. D. and L. J. Young (2005). Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**(5728): 1630-1634.

Hancock, J. M. (1999). Microsatellites and other simple sequences: genomic context and mutational mechanism. *Microsatellites: Evolution and Applications*. D. B. Goldstein and C. Schlotterer. New York, Oxford University Press.

Hancock, J. M. and J. S. Armstrong (1994). SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Computational Applications in Biosciences* **10**(1): 67-70.

Harbison, C. T., D. B. Gordon, et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99-104.

Harr, B. and C. Schlotterer (2000). Long microsatellite alleles in Drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* **155**(3): 1213-1220.

Harr, B., J. Todorova, et al. (2002). Mismatch repair-driven mutational bias in D. melanogaster. *Molecular Cell* **10**(1): 199-205.

Harr, B., B. Zangerl, et al. (2000). Removal of microsatellite interruptions by DNA replication slippage: Phylogenetic evidence from Drosophila. *Molecular Biology and Evolution* **17**(7): 1001-1009.

Hawk, J. D., L. Stefanovic, et al. (2005). Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *PNAS* **102**(24): 8639-8643.

Hazen, K. C. (1995). New and emerging yeast pathogens. *Clinical Microbiology Reviews* **8**(4): 462-78.

Henderson, S. T. and T. D. Petes (1992). Instability of Simple Sequence DNA in Saccharomyces-Cerevisiae. *Molecular and Cellular Biology* **12**(6): 2749-2757.

Herbert, A. and A. Rich (1996). The Biology of Left-handed Z-DNA. *Journal of Biological Chemistry.* **271**(20): 11595-11598.

Hinrichs, A. L., S. Bertelsen, et al. (2005). Multipoint identity-by-descent computations for single-point polymorphism and microsatellite maps. *BMC Genetics* **6**(SUPPL.1).

Hirsh, A. E., H. B. Fraser, et al. (2005). Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol Biol Evol* **22**(1): 174-7.

Huang, Q. Y., F. H. Xu, et al. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. *American Journal of Human Genetics* **70**(3): 625-634.

Huntley, M. A. and A. G. Clark (2007). Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution* **24**(12): 2598-2609.

Hussein, M. R., A. K. Haemel, et al. (2005). Genomic instability in radial growth phase melanoma cell lines after ultraviolet irradiation. *Journal of Clinical Pathology* **58**(4): 389-396.

Ihaka, R. and R. Gentleman (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**(3): 299-314.

International Mouse Genome Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Iyer, V. and K. Struhl (1995). Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *Embo Journal* **14**(11): 2570-9.

Jackson, A. L., R. Chen, et al. (1998). Induction of microsatellite instability by oxidative DNA damage. *Proceedings of the National Academy of Sciences* **95**(21): 12468-12473.

Jakupciak, J. P. and R. D. Wells (1999). Genetic instabilities in (CTG center dot CAG) occur by recombination. *Faseb Journal* **13**(7): A1543-A1543.

Jarne, P. and P. J. L. Lagoda (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**(10): 424-429.

Jeffreys, A. J., K. Tamaki, et al. (1994). Complex Gene Conversion Events in Germline Mutation at Human Minisatellites. *Nature Genetics* **6**(2): 136-145.

Jewell, E., A. Robinson, et al. (2006). SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery. *Nucleic Acids Res* **34** (*Web Server issue*): W656-9.

Jin, L., Y. X. Zhong, et al. (1994). The Exact Numbers of Possible Microsatellite Motifs. *American Journal of Human Genetics* **55**(3): 582-583.

Jurka, J., V. V. Kapitonov, et al. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**(1-4): 462-467.

Kalita, M., G. Ramasamy, et al. (2006). ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics* **7**(1): 336.

Karaoglu, H., C. M. Y. Lee, et al. (2005). Survey of simple sequence repeats in completed fungal genomes. *Molecular Biology and Evolution* **22**(3): 639-649.

Karlin, S., L. Brocchieri, et al. (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A* **99**(1): 333-8.

Kashi, Y., D. King, et al. (1997). Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics* **13**(2): 74-78.

Kashi, Y. and D. G. King (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* **22**(5): 253-259.

Katti, M. V., P. K. Ranjekar, et al. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution* **18**(7): 1161-1167.

Kayser, M., R. Kittler, et al. (2004). A comprehensive survey of human Y-chromosomal microsatellites. *American Journal of Human Genetics* **74**(6): 1183-1197.

Kayser, M., L. Roewer, et al. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *American Journal of Human Genetics* **66**(5): 1580-1588.

Kelkar, Y. D., S. Tyekucheva, et al. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research* **18**(1): 30-38.

Kellis, M., N. Patterson, et al. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**(6937): 241-254.

Kirkpatrick, D. T., Y. H. Wang, et al. (1999). Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. *Molecular and Cellular Biology* **19**(11): 7661-7671.

Kofler, R., C. Schlotterer, et al. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* **23**(13): 1683-5.

Kolpakov, R., G. Bana, et al. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**(13): 3672-3678.

Kroutil, L. C., K. Register, et al. (1996). Exonucleolytic proofreading during replication of repetitive DNA. *Biochemistry* **35**(3): 1046-1053.

Kruglyak, S., R. Durrett, et al. (2000). Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Molecular Biology and Evolution* **17**(8): 1210-1219.

Kruglyak, S., R. T. Durrett, et al. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America* **95**(18): 10774-10778.

Kurtz, S., J. V. Choudhuri, et al. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**(22): 4633-4642.

La Rota, M., R. V. Kantety, et al. (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *Bmc Genomics* **6**: -.

Lafyatis, R., F. Denhez, et al. (1991). Sequence specific protein binding to and activation of the TGF-ß3 promoter through a repeated TCCC motif. *Nucleic Acids Research* **19**(23): 6419-6425.

Lai, Y. L. and F. Z. Sun (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution* **20**(12): 2123-2131.

Landau, G. M., J. P. Schmidt, et al. (2001). An algorithm for approximate tandem repeats. *Journal of Computational Biology* **8**(1): 1-18.

Lander, E. S., L. M. Linton, et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Landry, P. A., M. T. Koskinen, et al. (2002). Deriving evolutionary relationships among populations using microsatellites and $\delta\mu^2$: all loci are equal, but some are more equal than others. *Genetics* **161**(3): 1339-47.

Leclercq, S., E. Rivals, et al. (2007). Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics* **8**(1): 125.

Legras, J. L., O. Ruh, et al. (2005). Selection of hypervariable microsatellite loci for the characterization of Saccharomyces cerevisiae strains. *International Journal of Food Microbiology* **102**(1): 73-83.

Levinson, G. and G. A. Gutman (1987b). Slipped-Strand Mispairing - a Major Mechanism for DNA-Sequence Evolution. *Molecular Biology and Evolution* **4**(3): 203-221.

Li, Y. C., A. B. Korol, et al. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11**(12): 2453-2465.

Li, Y. C., A. B. Korol, et al. (2004). Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution* **21**(6): 991-1007.

Lim, S., L. Notley-McRobb, et al. (2004). A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genetics and Biology* **41**(11): 1025-1036.

Liti, G. and E. J. Louis (2005). Yeast evolution and comparative genomics. *Annual Review of Microbiology* **59**: 135-153.

Liu, G., J. J. Bissler, et al. (2007). Unstable spinocerebellar ataxia type 10 (ATTCT)·(AGAAT) repeats are associated with aberrant replication at the *ATX10* locus and replication origin-dependent expansion at an ectopic site in human cells. *Molecular and Cellular Biology* **27**(22): 7828-7838.

Losa, R., S. Omari, et al. (1990). Poly(dA).Poly(dT) rich sequences are not sufficient to exclude nucleosome formation in a constitutive yeast promoter. *Nucleic Acids Research* **18**(12): 3495-3502.

Louis, E. J., E. S. Naumova, et al. (1994). The Chromosome End in Yeast - Its Mosaic Nature and Influence on Recombinational Dynamics. *Genetics* **136**(3): 789-802.

Lynch, M. and J. S. Conery (2003). The origins of genome complexity. *Science* **302**(5649): 1401-1404.

Lynch, M. and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, Ma., Sinauer.

Majewski, J. and J. Ott (2000). GT repeats are associated with recombination on human chromosome 22. *Genome Research* **10**(8): 1108-1114.

Malpertuy, A., B. Dujon, et al. (2003). Analysis of microsatellites in 13 hemiascomycetous yeast species: Mechanisms involved in genome dynamics. *Journal of Molecular Evolution* **56**(6): 730-741.

Marra, G. and P. Schar (1999). Recognition of DNA alterations by the mismatch repair system. *Biochemical Journal* **338**: 1-13.

Martin, A. P. and S. R. Palumbi (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America* **90**(9): 4087-91.

Martorell, L., K. Johnson, et al. (1997). Somatic instability of the myotonic dystrophy (CTG), repeat during human fetal development. *Human Molecular Genetics* **6**(6): 877-880.

Mattick, J. S. and I. V. Makunin (2006). Non-coding RNA. *Human Molecular Genetics* **15**(suppl_1): R17-29.

McCusker, J. H., K. V. Clemons, et al. (1994). Saccharomyces cerevisiae virulence phenotype as determined with CD-1 mice is associated with the ability to grow at 42 degrees C and form pseudohyphae. *Infectious Immunology* **62**(12): 5447-55.

Merkel, A. and N. Gemmell (2008). Detecting Microsatellites in Genome Data: Variance in Definitions and Bioinformatic Approaches Cause Systematic Bias. *Evolutionary Bioinformatics*(4): 1-6.

Meservy, J. L., R. G. Sargent, et al. (2003). Long CTG tracts from the myotonic dystrophy gene induce deletions and rearrangements during recombination at the APRT locus in CHO cells. *Molecular and Cellular Biology* **23**(9): 3152-62.

Messier, W., S. H. Li, et al. (1996). The birth of microsatellites. *Nature* **381**(6582): 483-483.

Metzgar, D., J. Bytof, et al. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research* **10**(1): 72-80.

Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature* **447**(7147): 932-940.

Moreira, J. M. A., J. E. Remacle, et al. (1998). Datin, a yeast poly(dA:dT)-binding protein, behaves as an activator of the wild-type ILV1 promoter and interacts synergistically with Reb1p. *Molecular and General Genetics* **258**(1-2): 95-103.

Morgante, M., M. Hanafey, et al. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**(2): 194-200.

Moxon, E. R., P. B. Rainey, et al. (1994). Adaptive Evolution of Highly Mutable Loci in Pathogenic Bacteria. *Current Biology* **4**(1): 24-33.

Moxon, E. R. and C. Wills (1999). DNA microsatellites: Agents of evolution? *Scientific American* **280**(1): 94-99.

Moxon, R., C. Bayliss, et al. (2006). Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics* **40**: 307-333.

Mudunuri, S. B. and H. A. Nagarajaram (2007). IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* **23**(10): 1181-7.

Mularoni, L., R. A. Veitia, et al. (2007). Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89**(3): 316-325.

Nadir, E., H. Margalit, et al. (1996). Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proceedings of the National Academy of Sciences of the United States of America* **93**(13): 6470-5.

Napierala, M., A. Bacolla, et al. (2005). Increased Negative Superhelical Density in Vivo Enhances the Genetic Instability of Triplet Repeat Sequences. *Journal of Biological Chemistry* **280**(45): 37366-37376.

Naslund, K., P. Saetre, et al. (2005). Genome-wide prediction of human VNTRs. *Genomics* **85**(1): 24-35.

Newlon, C. and J. Theis (2002). DNA replication joins the revolution: Whole-genome views of DNA replication in budding yeast. *BioEssays* **24**(4): 300-304.

Nieduszynski, C. A., Y. Knox, et al. (2006). Genome-wide identification of replication origins in yeast by comparative genomics. *Genes and Development* **20**(14): 1874-1879.

Nishant, K. T. and M. R. S. Rao (2006). Molecular features of meiotic recombination hot spots. *Bioessays* **28**(1): 45-56.

O'Dushlaine, C. T., R. J. Edwards, et al. (2005). Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biology* **6**(8): R69.

O'Dushlaine, C. T. and D. C. Shields (2006). Tools for the identification of variable and potentially variable tandem repeats. *BMC Genomics* **7**: 290.

Ohta, T. and M. Kimura (1973). New Model for Estimating Number of Electrophoretically Detectable Alleles in a Finite Population. *Genetics* **74**(JUN): S201-S201.

Okladnova, O., Y. V. Syagailo, et al. (1998). A promoter-associated polymorphic repeat modulates PAX-6 expression in human brain. *Biochemistry and Biophysics Research Communications* **248**(2): 402-5.

Page, R. and E. Holmes (1998). *Molecular Evolution: A Phylogenetic Approach*. Oxford, Blackwell Science.

Pardi, F., R. M. Sibly, et al. (2005). On the structural differences between markers and genomic AC microsatellites. *Journal of Molecular Evolution* **60**(5): 688-93.

Pavlov, Y. I., I. M. Mian, et al. (2003). Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. *Current Biology* **13**(9): 744-748.

Payseur, B. A. and M. W. Nachman (2000). Microsatellite variation and recombination rate in the human genome. *Genetics* **156**(3): 1285-1298.

Pearson, C. E., K. N. Edamura, et al. (2005). Repeat instability: Mechanisms of dynamic mutations. *Nature Reviews Genetics* **6**(10): 729-742.

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics* **2**(5): 360-9.

Petes, T. D., P. W. Greenwell, et al. (1997). Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. *Genetics* **146**(2): 491-498.

Primmer, C. R. and H. Ellegren (1998). Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution* **15**(8): 997-1008.

Primmer, C. R., H. Ellegren, et al. (1996). Directional evolution in germline microsatellite mutations. *Nature Genetics* **13**(4): 391-393.

Primmer, C. R., N. Saino, et al. (1998). Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallows Hirundo rustica. *Molecular Biology and Evolution* **15**(8): 1047-1054.

Quinn, G. and M. Keough (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge, Cambridge University Press.

Ramsay, L., M. Macaulay, et al. (1999). Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant Journal* **17**(4): 415-425.

Richard, G. F. and B. Dujon (2001). Stability of CAG trinucleotide repeats during I-Sce I induced meiotic recombination in yeast. *Yeast* **18**: S41-S41.

Richard, G. F., C. Hennequin, et al. (1999). Trinucleotide repeats and other microsatellites in yeasts. *Research in Microbiology* **150**(9-10): 589-602.

Richard, G. F., C. Cyncynatus, et al. (2003). Contractions and expansions of CAG/CTG trinucleotide repeats occur during ectopic gene conversion in yeast, by a MUS81-independent mechanism. *Journal of Molecular Biology* **326**(3): 769-782.

Rolfsmeier, M. L., M. J. Dixon, et al. (2000). Mismatch repair blocks expansions of interrupted trinucleotide repeats in yeast. *Molecular Cell* **6**(6): 1501-1507.

Rolfsmeier, M. L. and R. S. Lahue (2000). Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **20**(1): 173-180.

Rose, O. and D. Falush (1998). A threshold size for microsatellite expansion. *Molecular Biology and Evolution* **15**(5): 613-615.

Ross, C. L., K. A. Dyer, et al. (2003). Rapid divergence of microsatellite abundance among species of Drosophila. *Molecular Biology and Evolution* **20**(7): 1143-1157.

Rozen, S. and H. Skaletsky (2000). Primer3 on the WWW for general user and for biologists programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. S. Krawetz and S. Misener. Totowa, NJ, Humana Press**:** 365-386.

Ruitberg, C. M., D. J. Reeder, et al. (2001). STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research* **29**(1): 320-322.

Sainudiin, R., R. T. Durrett, et al. (2004). Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. *Genetics* **168**(1): 383-395.

Santibanez-Koref, M. F., R. Gangeswaran, et al. (2001). A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Molecular Biology and Evolution* **18**(11): 2119-2123.

Satchwell, S. C., H. R. Drew, et al. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191**(4): 659-75.

Sawyer, L. A., J. M. Hennessy, et al. (1997). Natural variation in a Drosophila clock gene and temperature compensation. *Science* **278**(5346): 2117-2120.

Schlotterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**(6): 365-371.

Schlotterer, C. (2003). Hitchhiking mapping--functional genomics from the population genetics perspective. *Trends in Genetics* **19**(1): 32-8.

Schlotterer, C. (2004). The evolution of molecular markers - Just a matter of fashion? *Nature Reviews Genetics* **5**(1): 63-69.

Schlotterer, C., R. Ritter, et al. (1998). High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution* **15**(10): 1269-1274.

Schug, M. D., C. M. Hutter, et al. (1998). The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster. *Molecular Biology and Evolution* **15**(12): 1751-1760.

Schuler, G. D. (1998). Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends in Biotechnology* **16**(11): 456-9.

Schultes, N. P. and J. W. Szostak (1991). A Poly(dA-dT) tract is a component of the recombination initiation site at the Arg4 locus in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **11**(1): 322-328.

Segal, E., Y. Fondufe-Mittendorf, et al. (2006). A genomic code for nucleosome positioning. *Nature* **442**(7104): 772-8.

Selkoe, K. A. and R. J. Toonen (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**(5): 615-629.

Shanker, A., A. Singh, et al. (2007). In silico mining in expressed sequences of Neurospora crassa for identification and abundance of microsatellites. *Microbiology Research* **162**(3): 250-6.

Sharma, P. C., A. Grover, et al. (2007). Mining microsatellites in eukaryotic genomes. *Trends in Biotechnology* **25**(11): 490-8.

Sia, E. A., S. JinksRobertson, et al. (1997). Genetic control of microsatellite stability. *Mutation Research-DNA Repair* **383**(1): 61-70.

Sia, E. A., R. J. Kokoska, et al. (1997). Microsatellite instability in yeast: Dependence on repeat unit size and DNA mismatch repair genes. *Molecular and Cellular Biology* **17**(5): 2851-2858.

Smit, A. F. A. and P. Green (1996). RepeatMasker. http://www.repeatmasker.org/

Smith, J. M. and J. Haigh (1974). The hitch-hiking effect of a favourable gene. *Genetic Research* **23**(1): 23-35.

Smith, T. F. and M. S. Waterman (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* **147**(1): 195-7.

Spencer, C. C., P. Deloukas, et al. (2006). The influence of recombination on human genetic diversity. *PLoS Genetics* **2**(9): e148.

Strand, M., T. A. Prolla, et al. (1993). Destabilization of Tracts of Simple Repetitive DNA in Yeast by Mutations Affecting DNA Mismatch Repair. *Nature* **365**(6443): 274-276.

Subramanian, S., R. K. Mishra, et al. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology* **4**(2).

Suter, B., D. Auerbach, et al. (2006). Yeast-based functional genomics and proteomics technologies: The first 15 years and beyond. *BioTechniques* **40**(5): 625-644.

Suter, B., G. Schnappauf, et al. (2000). Poly(dA center dot dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Research* **28**(21): 4083-4089.

Taylor, J. S., J. M. H. Durkin, et al. (1999). The death of a microsatellite: A phylogenetic perspective on microsatellite interruptions. *Molecular Biology and Evolution* **16**(4): 567-572.

Temnykh, S., G. DeClerck, et al. (2001). Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* **11**(8): 1441-1452.

Temnykh, S., W. D. Park, et al. (2000). Mapping and genome organization of microsatellite sequences in rice (Oryza sativa L.). *Theoretical and Applied Genetics* **100**(5): 697-712.

Thiel, T., W. Michalek, et al. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics* **106**(3): 411-422.

Thurston, M. I. and D. Field (2005). Msatfinder: detection and characterization of microsatellites. CEH Oxford.

Toth, G., Z. Gaspari, et al. (2000). Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research* **10**(7): 967-981.

Trivedi, S. (2006). Comparison of simple sequence repeats in 19 Archaea. *Genetics and Molecular Research* **5**(4): 741-772.

Urquhart, A., C. P. Kimpton, et al. (1994). Variation in Short Tandem Repeat sequences a survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine* **107**(1): 13-20.

Valdes, A. M., M. Slatkin, et al. (1993). Allele Frequencies at Microsatellite Loci - the Stepwise Mutation Model Revisited. *Genetics* **133**(3): 737-749.

Valle, G. (1993). Ta-Repeat Microsatellites Are Closely Associated with Ars Consensus Sequences in Yeast Chromosome-Iii. *Yeast* **9**(7): 753-759.

van Belkum, A., S. Scherer, et al. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and Molecular Biology Reviews* **62**(2): 275-293.

Vargas Jentzsch, I., A. Bagshaw, et al. (2008). Evolution of Microsatellite DNA. *Encyclopedia of Life Sciences*. Chichester http://www.els.net/, John Wiley & Sons.

Verstrepen, K. J., A. Jansen, et al. (2005). Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**(9): 986-990.

Verstrepen, K. J., T. B. Reynolds, et al. (2004). Origins of variation in the fungal cell surface. *Nature Reviews Microbiology* **2**(7): 533-40.

Wang, Y. H., R. Gellibolian, et al. (1996). Long CCG triplet repeat blocks exclude nucleosomes: A possible mechanism for the nature of fragile sites in chromosomes. *Journal of Molecular Biology* **263**(4): 511-516.

Wang, Y. H. and J. D. Griffith (1996). The $[(G/C)_3NN]_n$ motif: A common DNA repeat that excludes nucleosomes. *Proceedings of the National Academy of Sciences of the United States of America* **93**(17): 8863-8867.

Warren, W. C., L. W. Hillier, et al. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**(7192): 175-83.

Wells, R. D., R. Dere, et al. (2005). Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Research* **33**(12): 3785-98.

Wexler, Y., Z. Yakhini, et al. (2005). Finding approximate tandem repeats in genomic sequences. *Journal of Computational Biology* **12**(7): 928-942.

Widom, J. (2001). Role of DNA sequence in nucleosome stability and dynamics. *Quarterly Reviews Biophysics* **34**(3): 269-324.

Wierdl, M., M. Dominska, et al. (1997). Microsatellite instability in yeast: Dependence on the length of the microsatellite. *Genetics* **146**(3): 769-779.

Wierdl, M., C. N. Greene, et al. (1996). Destabilization of simple repetitive DNA sequences by transcription in yeast. *Genetics* **143**(2): 713-721.

Wilder, J. and H. Hollocher (2001). Mobile elements and the genesis of microsatellites in dipterans. *Molecular Biology and Evolution* **18**(3): 384-392.

Wren, J. D., E. Forgacs, et al. (2000). Repeat polymorphisms within gene regions: Phenotypic and evolutionary implications. *American Journal of Human Genetics* **67**(2): 345-356.

Xu, G. and A. G. Goodridge (1998). A CT repeat in the promoter of the chicken malic enzyme gene is essential for function at an alternative transcription start site. *Archives of Biochemistry and Biophysics* **358**(1): 83-91.

Xu, H., R. Chakraborty, et al. (2005). Mutation rate variation at human dinucleotide microsatellites. *Genetics* **170**(1): 305-12.

Xu, X., M. Peng, et al. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics* **24**(4): 396-399.

Yon Rhee, S., V. Wood, et al. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics* **9**(7): 509-515.

Young, E. T., J. S. Sloan, et al. (2000). Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. *Genetics* **154**(3): 1053-1068.

Zhang, N., A. L. Harrex, et al. (2003). Sixty alleles of the ALS7 open reading frame in Candida albicans: ALS7 is a hypermutable contingency locus. *Genome Research* **13**(9): 2005-17.

Zhu, Y., J. E. Strassmann, et al. (2000). Insertions, substitutions, and the origin of microsatellites. *Genetical Research* **76**(3): 227-236.

Webpage references

Expasy tools. http://www.expasy.ch/tools.html (04/2008)

Galaxy web-server. http://g2.trac.bx.psu.edu/; (2007, 2008)

GC content. http://tim.saraogtim.com/molbio/gccontent.php; (03/2008)

Saccharomyces Genome Database (SGD). http://www.yeastgenome.org/.(2007)

Saccharomyces Genome Resequencing Project (SGRP). http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml; (11/2007).

R statistical software. http://cran.r-project.org/; (2006)

Appendix

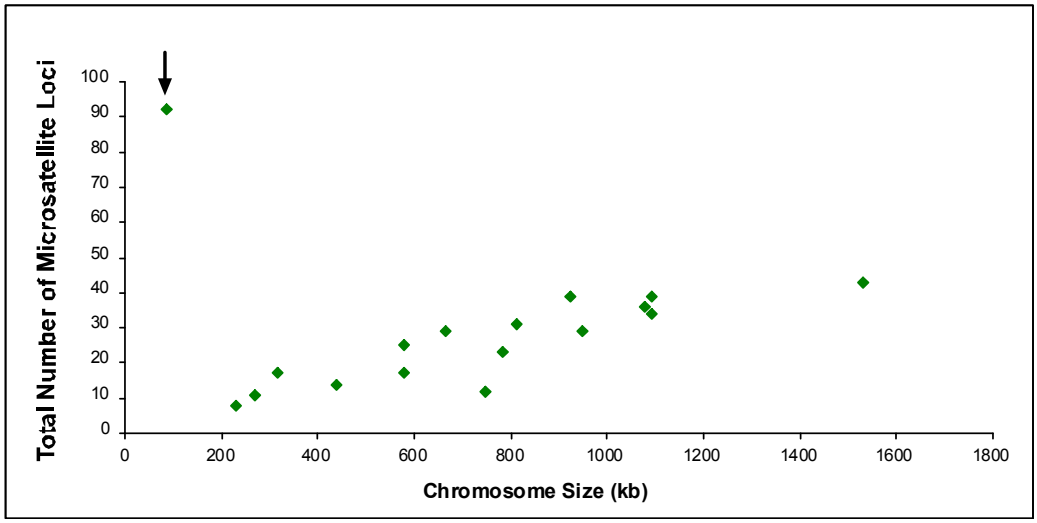
A.1. Additional Figures

A.1.2. Chapter 2

Table 1. Variation in TRF results* between genome builts

Date Genome Built	1/01/1998	1/10/2003	30/11/2006
Total Sequence Size (nuclear), nt	12069303	12070521	12070899
Repeats found with TRF (default)	406	407	406

Figure 1. Variation in microsatellite abundance between different chromosome and mtDNA (↓). Note the roughly linear relationship between loci number and chromosome size with mtDNA (↓) as outlier.



Sequences were downloaded from ftp at SGD (ftp://genome-ftp.stanford.edu/pub/yeast/sequence/ NCBI_genome_source); *TRF default parameters: 2 7 7 80 10 50 6 (minimum length: 25nt)

A.1.3. Chapter 4

Table 1. Distribution and characteristics of repeat unit sizes. For comparison results are shown for two different sets of minimum length threshold in bp (repeat length): i) staggered minimum length thresholds, and ii) equal minimum array length for short (mono-trinucleotides) and long (tetra-hexanucleotides) repeat sizes.

i) Perfect repeats - minimum array length: 12(1), 12(2), 12(3), 16(4), 20(5), 24(6)

Motif	Counts	Average_Length	Counts/Mbp	GC-Content	StdDeviation_AverageLength
mononucleotide	1120	14.92	92.79	0.00	4.03
dinucleotide	449	18.14	37.20	0.08	6.88
trinucleotide	1020	15.91	84.50	0.33	7.04
tetranucleotide	63	19.25	5.22	0.14	5.35
pentanucleotide	29	23.17	2.40	0.27	3.69
hexanucleotide	51	29.18	4.23	0.44	5.26

ii) Perfect repeats- minimum array length: 12(1), 12(2), 12(3), 16(4), 15(5), 18(6)

Motif	Counts	Average_Length	Counts/Mbp	GC-Content	StdDeviation_AverageLength
Pentanucleotide*	152	17.38	12.59	0.30	3.41
Hexanucleotide*	255	21.58	21.13	0.44	4.71

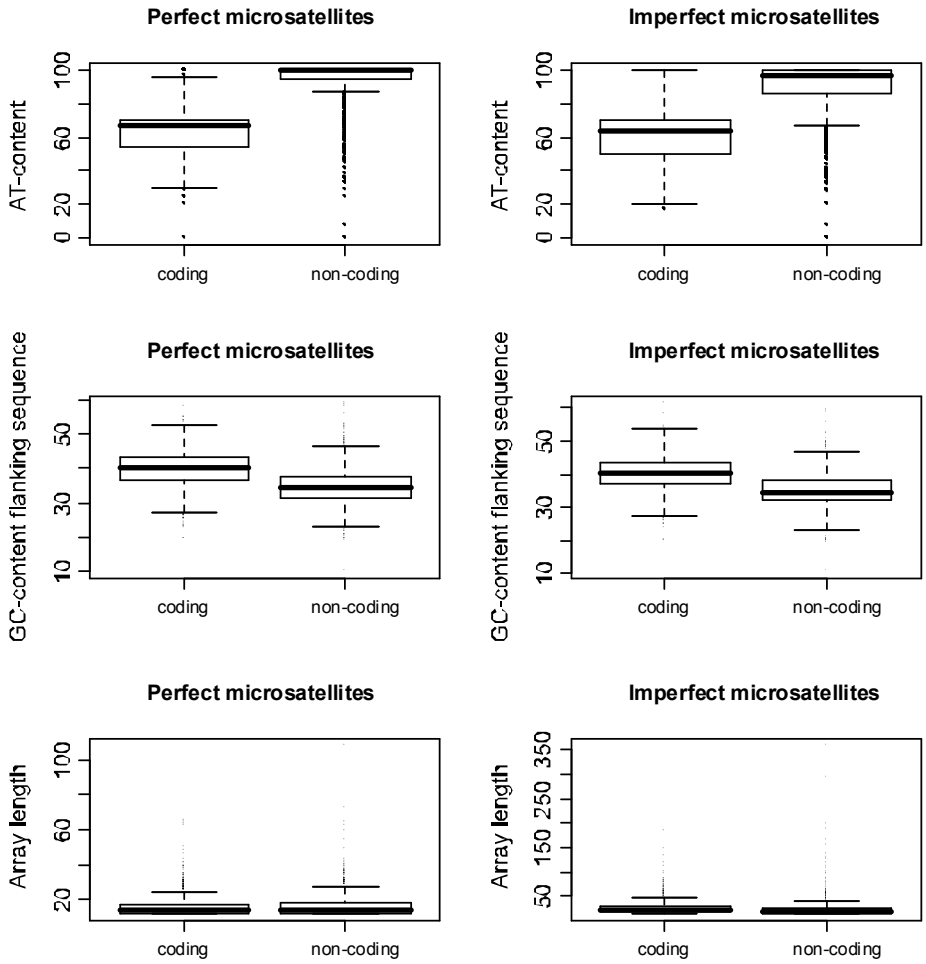


Figure 1: Distribution of perfect and imperfect (14n4) microsatellites in coding and non-coding regions (95% CI shown). Microsatellites in coding regions show a strong GC-enrichment. Imperfect microsatellites in non-coding regions are less AT-rich than perfect microsatellites. There is no difference in array length between genomic regions.

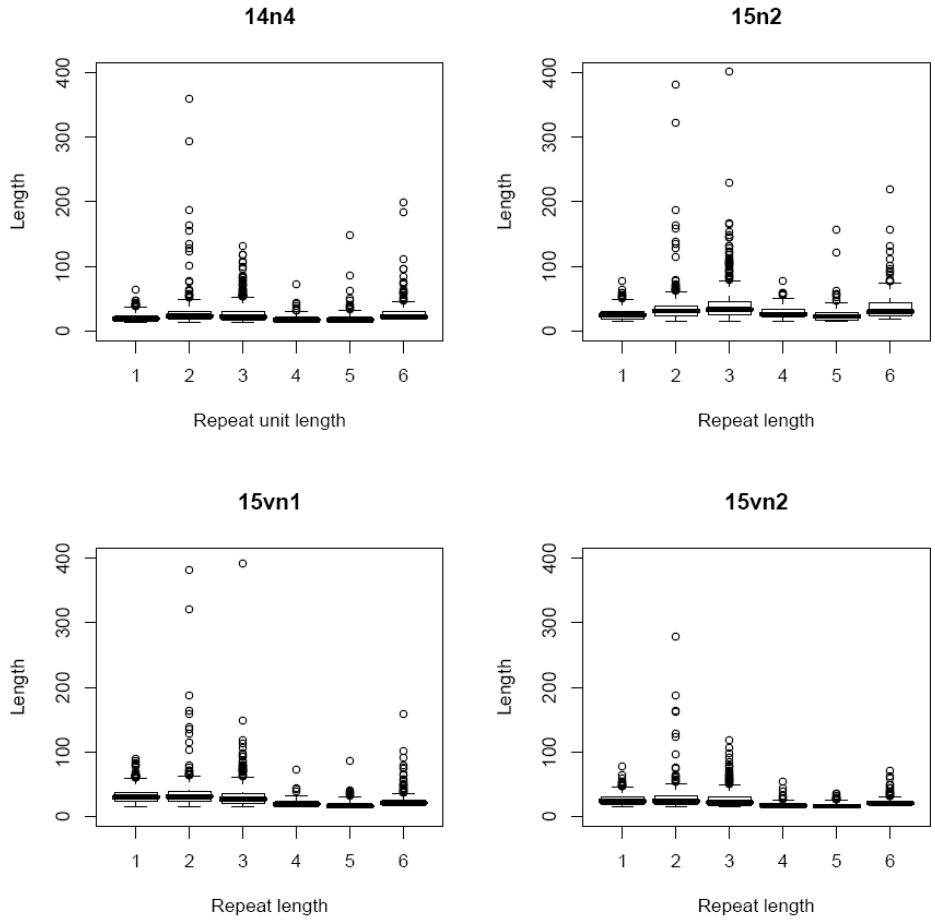


Figure 2. Influence of parameter settings on array length for different size classes**.** Under decreased variable mismatch penalty (15n2), there is a increase in array length for tri- and hexanucleotides; whereas a lower fixed mismatch (15vn1) penalty increases array length predominantly in smaller unit sizes.

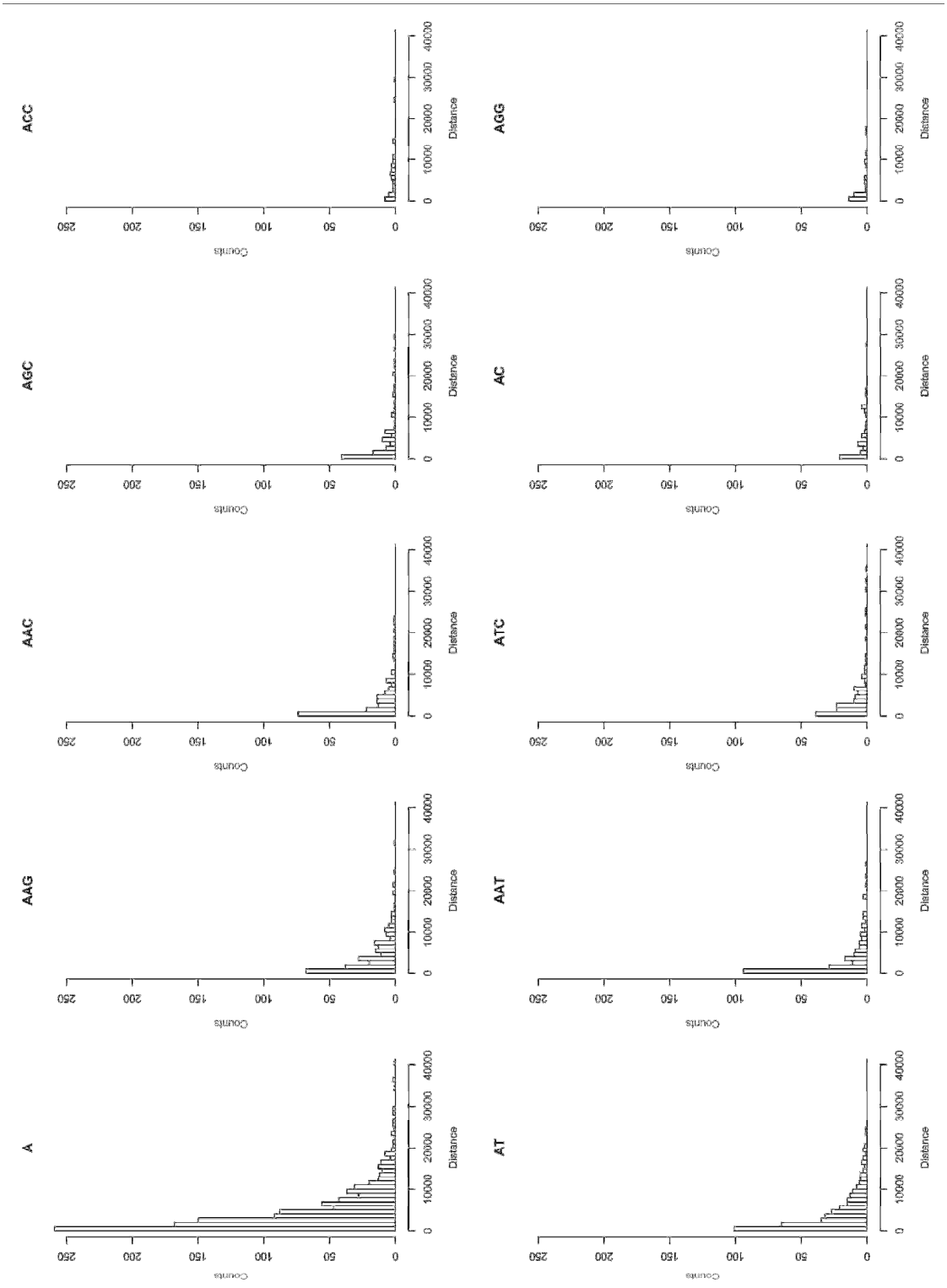


Figure 3: Distance between adjacent microsatellite loci in the yeast genome (show are only the 10 most commonly found motifs)

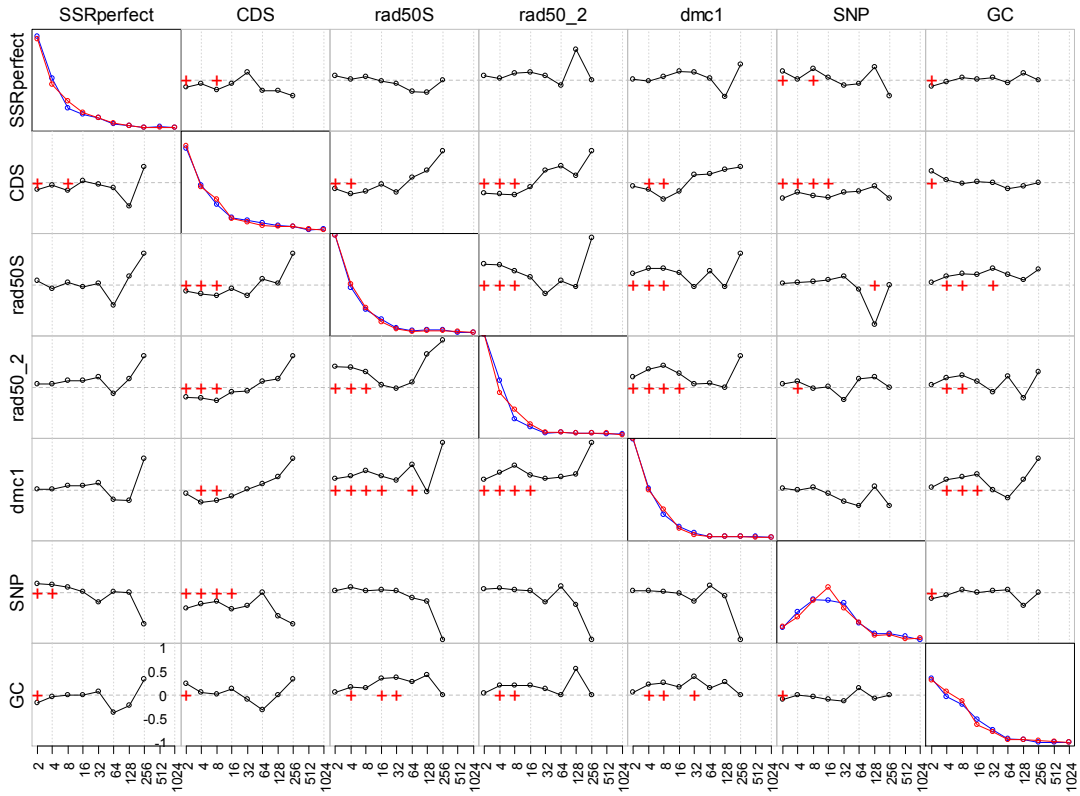


Figure 4: Example for pairwise correlations of wavelet coefficients in yeast chromosomes (chromosome 15), Lower triangle shows the chromosome start region, the upper triangle shows the chromosome end region. Red crosses indicate significant correlation (Kendall's rank correlation, p-value <0.01). Scale is given in kilobases. Diagonal plots depict power spectra for each factor. (The power distribution of a signal describes the proportions of the total variance explained by the heterogeneity at different scales, hence is measurement for inconsistency in the signal)

A.1.4. Chapter 5

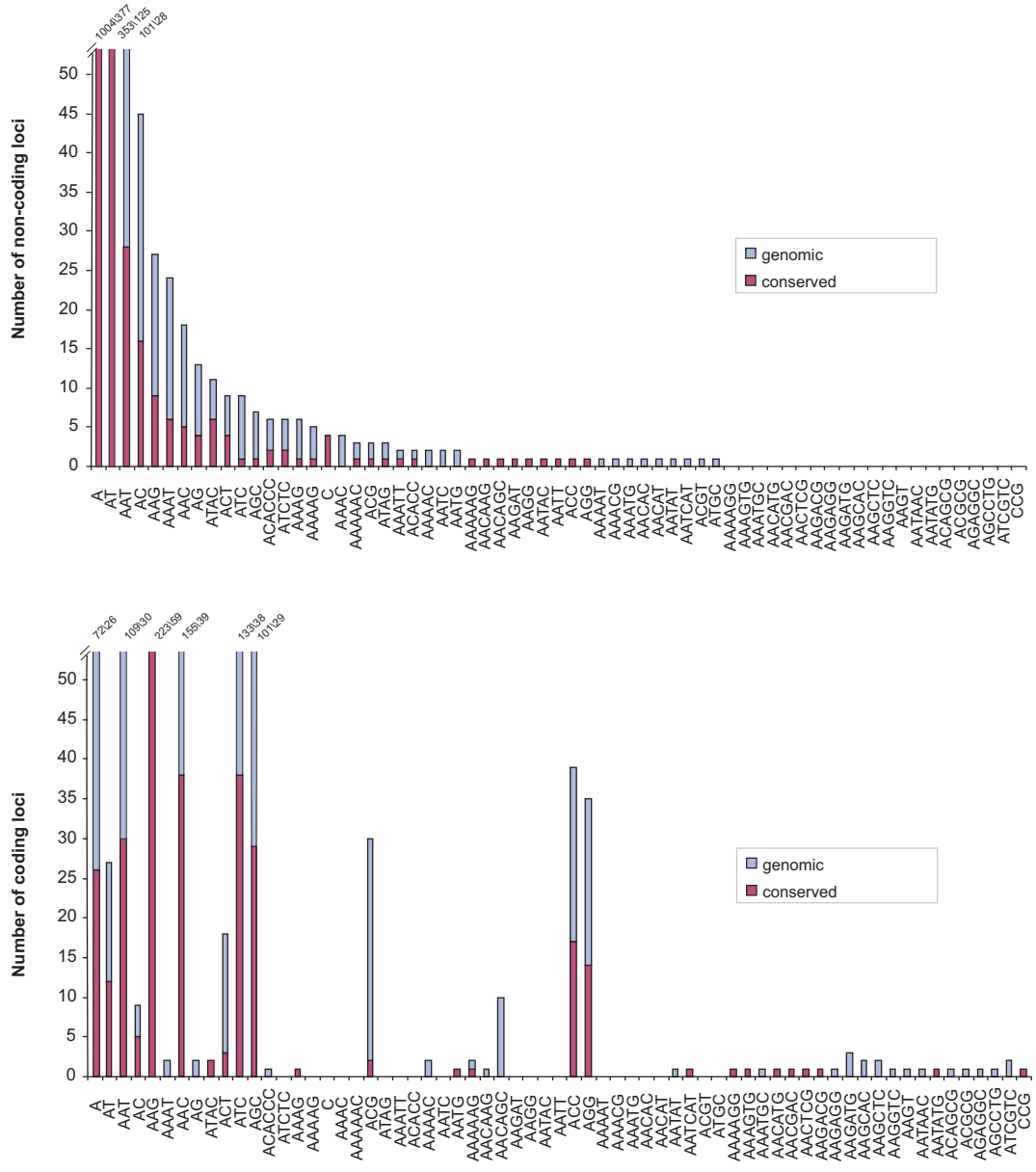


Figure 1. Distribution of motifs in a) the total genomic population of microsatellite loci, and b) conserved microsatellite loci, across coding and non-coding regions. [Not all motifs are conserved purely as a result of their genomic frequency. For example, poly(C), a motif which is considered rare within eukaryotic genomes, is conserved in all four cases of its occurrence within the non-coding fraction of the genome (one of which also shows weak length polymorphism.). On the other hand, within coding regions, motifs that have moderate frequencies like ACT_n (18 genomic, 3 conserved loci) and ACG_n (30 genomic, 2 conserved loci) are barely conserved at all.]

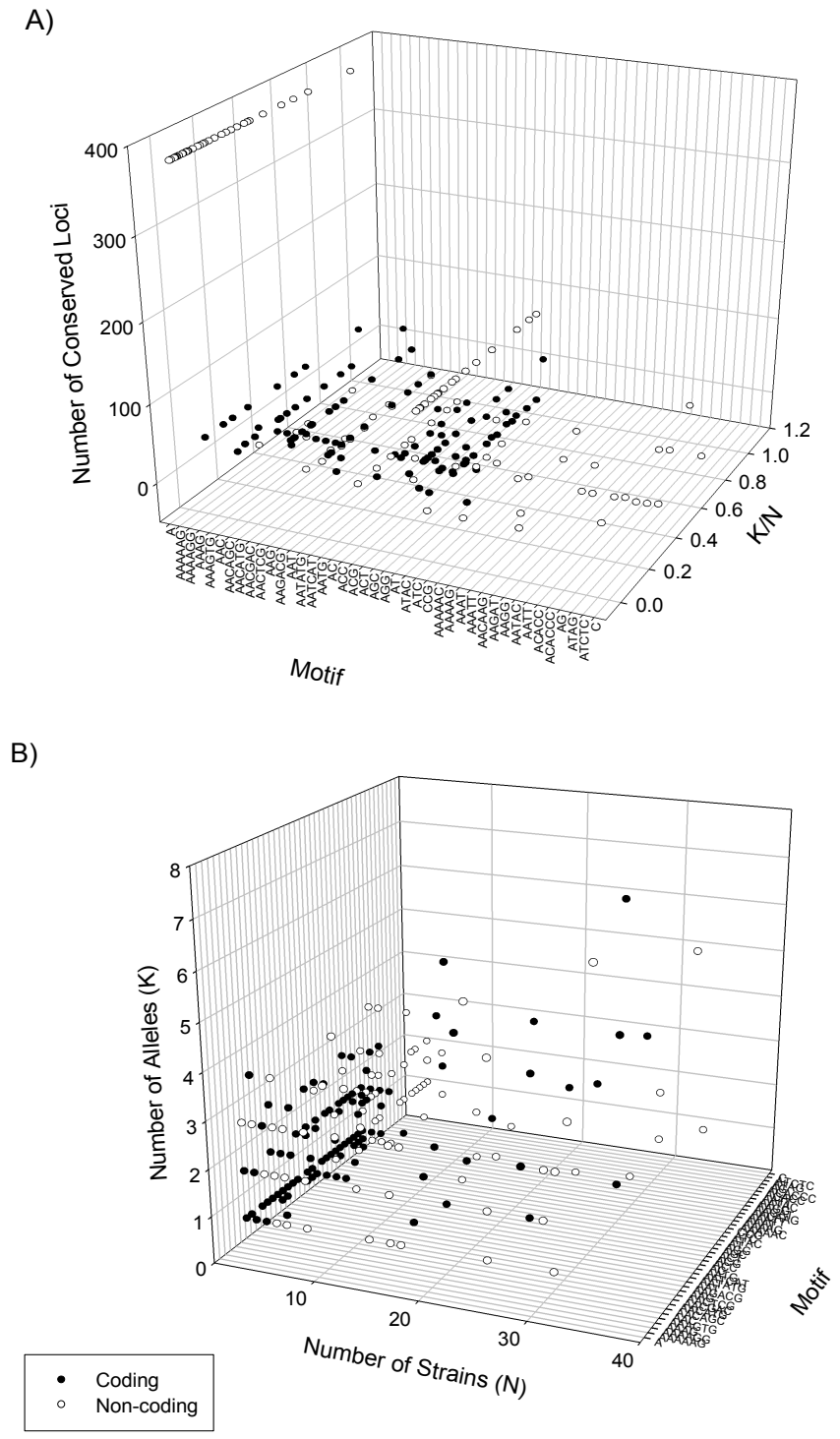


Figure 2. Conservation and polymorphism of microsatellite loci across different motifs type for coding and non-coding fractions

SSR 339

Motif: ACA (19nt) GACAACAACAACAACAACA QQQQQ

RefLocation: Chr3: 283214 – 28323bp (YCR093W: 380114 – 286440bp)

A)

```
SGD_Scer_CDC39/YCR093W  3061  ATGCAACAACACCAGCAACAGATGCTAATATATCAACAGAGACAACAACA  3110
MIT_Sbay_c609_2996      3115  ATGCAACAACACCAACAGCAAATGCTAATATATCAACAAAGACAACAACA  3164
MIT_Smik_c295_2289      3061  ATGCAACAACATCAACAACAGATGCTATTATATCAGCAAAGGCAACAGCA  3110
MIT_Spar_c90_2756       3061  ATGCAACAGCACCAACAGCAGATGCTAATATACCAACAGAGACAGCAACA  3110
WashU_Sbay_Contig666.4  3115  ATGCAACAACACCAACAGCAAATGCTAATATATCAACAAAGACAACAACA  3164
WashU_Scas_Contig720.101 3061 TTACAACAACATCAACAACAAATGATGATTCTCCAACAAAGACAACAACG  3110
Symbols                       * ***** ** ** ** ** *** *  *    ** ** ** ** ** *

SGD_Scer_CDC39/YCR093W  3122  GGCAACAACAACAACAACATCATATTAGTGCAAATACAATCGCAGACCAA  3171
MIT_Sbay_c609_2996      3215  AGCAACAACAACAACATCATCATATGGGTACAAATCCTGTCGCTGATCAA  3264
MIT_Smik_c295_2289      3118  --CAAAGACAACAACAACATCATATGAGTGCGAACACAATTTCAGATCAG  3165
MIT_Spar_c90_2756       3125  AACAACAACAACAACAGCATCATATGAGTGCGAATACAATCACAGATCAA  3174
WashU_Sbay_Contig666.4  3215  AGCAACAACAACAACATCATCATATGGGTACAAATCCTGTCGCTGATCAA  3264
WashU_Scas_Contig720.101 3113 ------------------------TGGTTTCTGGAGCAATATCAGA---A  3135
Symbols                                              *   * *     *  *  * **

SGD_Scer_CDC39/YCR093W  3111  ACAACAA----------------------------------------CAAA  3121
MIT_Sbay_c609_2996      3165  ACAACAACAGCAACAACAACAACAGCAACAGCAACAGCAACAGCAACAAC  3214
MIT_Smik_c295_2289      3111  GCAGCAA----------------------------------------    3117
MIT_Spar_c90_2756       3111  ACAACAAAGG--------------------------------------CAAC  3124
WashU_Sbay_Contig666.4  3165  ACAACAACAGCAACAACAACAACAGCAACAGCAACAGCAACAGCAACAAC  3214
WashU_Scas_Contig720.101 3111 -----TA----------------------------------------    3112
Symbols                       *
```

B)

```
SGD_Scer_CDC39/YCR093W   1024  HQQQMLIYQQRQQQQQ----------QRQQQQQHHISANTIADQQAA  1060
MIT_Sbay_c609_2996       1042  HQQQMLIYQQRQQQQQQQQQQQQQQQQQQQQQQQQHHHMGTNPVADQQAT  1091
MIT_Smik_c295_2289       1024  HQQQMLLYQQRQQQQQ----------QRQQQHHMSANTISDQQTT  1058
MIT_Spar_c90_2756        1024  HQQQMLIYQQRQQQQQR----------QQQQQQQHHMSANTITDQQTN  1061
WashU_Sbay_Contig666.4   1042  HQQQMLIYQQRQQQQQQQQQQQQQQQQQQQQQQQQHHHMGTNPVADQQAT  1091
WashU_Scas_Contig720.101 1024  HQQQMMILQQRQQRMVSG----------AISEQVP  1048
Symbols                        *****:: *****:                .:*.
```

Figure 3. DNA (A) and amino acid (B) sequence alignment of a conserved microsatellite locus across the *sensu strictu* (*Saccharomyces*) group. Scer = *S.cerevisiae*, Sbay = *S. bayanus*, Smik = *S. mikatae*, Spar = *S.paradoxus*, Scas = *S.castellii*. (Fungal BLAST with WU-BLAST2 at SGD, http://seq.yeastgenome.org/cgi-bin/blast-fungal.pl, using default parameters)

A.2. Published Papers

Merkel, A. and N. Gemmell (2008). Detecting microsatellites in genome data: variance in definitions and bioinformatic approaches cause systematic bias. *Evolutionary Bioinformatics* (4): 1-6.

Merkel, A. and N. Gemmell (2008). Detecting short tandem repeats from genome data: opening the software black box. *Briefings in Bioinformatics* **9**(5): 355-366

Vargas Jentzsch, I., A. Bagshaw, et al. (2008). Evolution of Microsatellite DNA. *Encyclopedia of Life Sciences*. Chichester http://www.els.net/, John Wiley & Sons.

Publications not included in this thesis:

Warren, W. C., L. W. Hillier, et al. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**(7192): 175-83.

Detecting Microsatellites in Genome Data: Variance in Definitions and Bioinformatic Approaches Cause Systematic Bias

Angelika Merkel and Neil J. Gemmell

School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.

Abstract: Microsatellites are currently one of the most commonly used genetic markers. The application of bioinformatic tools has become common practice in the study of these short tandem repeats (STR). However, *in silico* studies can suffer from study bias. Using a meta-analysis on microsatellite distribution in yeast we show that estimates of numbers of repeats reported by different studies can differ in the order of several magnitudes, even within a single genome. These differences arise because varying definitions of microsatellites, spanning repeat size, array length and array composition, are used in different search paradigms, with minimum array length being the main influencing factor. Structural differences in the implemented search algorithm additionally contribute to variation in the number of repeats detected. We suggest that for future studies a consistent approach to STR searches is adopted in order to improve the power of intra- and interspecific comparisons

Keywords: microsatellites, short tandem repeats, definition, genome, array length, study bias

Introduction

Microsatellites or short sequence/tandem repeats (SSRs/ STRs) are tandemly repeated DNA sequences of (commonly) 1–6bp length per repeat unit. Their high length polymorphism and abundance in all genomes make them the genetic marker of choice for a diverse range of applications spanning linkage analysis and genetic mapping through to forensics and ecological and evolutionary studies (Goldstein and Schlötterer, 1999). Interest in microsatellite mutational dynamics is increasing, with significant interest emerging in the use of genomic data to investigate the evolution of these ubiquitous and useful sequences. To date, a significant number of studies have investigated microsatellite abundance in a range of species in order to examine the evolution of these simple sequences and infer their functional roles, if any, in gene regulation, genome structure etc. (Kashi and King, 2006). Putative distribution biases have been investigated for introns, exons and intergenic regions as well as possible associations with other genomic elements, such as interspersed repeats (Arcot et al. 1995; Li et al. 2004; Lim et al. 2004; Malpertuy et al. 2003; Toth et al. 2000).

However, comparisons among large scale *in silico* genome studies, even from the same genomic data, are fraught with methodological bias. A recent paper by Leclercq et al. (2007) outlines significant differences among search algorithms based on intrinsic structure of the search algorithm and the parameter settings. We present a meta-analysis on microsatellite distribution in yeast as an example on how divergent study results can be in practice. We confirm Leclercq's (2007) findings, but more importantly we show that the differences are rooted in a long-lived controversy, ever since microsatellites were first discovered 20 years ago; how exactly to define a microsatellite. Interspecies comparisons that derive from different studies are particularly vulnerable to erroneous conclusions, and it is an intricate task to tease out the patterns of microsatellite evolution from those arising from study bias.

Methods

We undertook a meta-analysis of the published literature on microsatellite distribution in the yeast genome (*Saccharomyces cerevisiae*). The studies chosen are all comparisons of microsatellite distribution

Correspondence: Angelika Merkel, School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. Tel: +64 (0) 3 364 2987 x7048; Fax: +64 (0) 3 364 2590; Email: ame52@student.canterbury.ac.nz

patterns (motif, size class, and array length) that include *S. cerevisiae* as one of the focal species, but differ in the approach and software used to detect microsatellite sequences (Table 1).

Results

All analyzed studies confirm unique species-specific motif distribution patterns and an over-representation of long arrays over short arrays, which is in concordance with current models of microsatellite evolution. However, we find striking differences in the reported results (Figure 1). For example, Dieringer and Schlotterer, (2003) report more repeats across all motif types than others, up to several magnitudes difference. This study scored repeat frequencies (loci/Mbp) in the order of 104 for di- and trinucleotides and 103 for tetranucleotides, compared to 102 for dinucleotides and 101 for tri-and tetranucleotides, which are the next highest frequencies out of all other studies. Among all repeat sizes, mononucleotides are especially variable in the numbers of loci reported. We found frequency counts that ranged from a minimum of 46 loci/Mbp (Katti, Ranjekar, and Gupta, 2001) to a maximum of 142,200 loci/Mbp (Dieringer and Schlotterer, 2003). The relative abundance of size classes also differs among studies. For example, all studies report mononucleotides as the most abundant size class with decreasing frequencies of longer repeat units, except Katti et al. (2001) who report the highest numbers for trinucleotides and van Belkum et al. (1998) who show an increased frequency for penta- and hexanucleotides.

Discussion

Given that the seven studies we examined have essentially analyzed the same genome data (small variations in build version not withstanding) for the

Table 1. Studies utilized in the meta-analysis. All studies report comparisons of microsatellite distribution pattern in yeast. Table shows (from left to right) study, algorithm or software employed, the type of repeat that was investigated (with respect to perfection/imperfection) and parameter that were implemented in the bioinformatics search, such as repeat size (mono-octanucleotide) and array length (minimum/maximum threshold).

Study	Algorithm	Type of repeat	Repeat parameters
Field and Wills (1998)	PERL script –regular expression[1]	Perfect repeats	All mononucleotides: 1–42bp Repeat size: 2, 3, 4, 5, 6bp Minimum length: 16, 24, 32, 40, 48, 56, 64bp
van Belkum et al. (1998)	C-script[2]	Perfect repeats	Repeat size: 1, 2, 3, 4, 5, 6, 7, 8bp Minimum length: 10, 10, 18, 20, 18, 20, 21, 24bp
Katti Ranjekar and Gupta (2001)	C-script, –base-by-base search using adjacent sliding windows for alignments	Imperfect repeats (mismatch every 10th nt)	Repeat size: 1, 2, 3, 4bp Minimum length: 20, 20, 21, 20bp
Dieringer and Schlötterer (2003)	C-script, –motif search for consecutive sequence stretches	Perfect repeats (incl. partial copies)	Repeat size: 1, 2, 3, 4bp Minimum length: 2, 4, 6, 8bp Maximum length: 20bp
Malpertuy, Dujon and Richard (2003)	TRF software (Benson 1999), –statistic/ heuristic approach	Imperfect repeats (match: (+1) mismatch: (–2, –3, –4) indels: (–6, –9, –12))	Pattern size: 2, 3, 4bp Minimum length: 10, 15, 20bp Maximum length: 20 repeats
Karaoglu, Lee and Meyer (2005)	PYTHON script	Perfect repeats	Pattern size: 1, 2, 3, 4, 5, 6bp Minimum length: 10bp
Lim et al. (2004)	C++ script, –base-by-base search using adjacent sliding windows for alignment	Perfect repeats	Pattern size: 1, 2, 3, 4, 5, 6bp Minimum length: 5 repeats

[1]Personal communication, algorithm is now implemented as *MsatFinder* software (http://www.bioinf.ceh.ac.uk/msatfinder/).
[2]The URL address given for the server was not valid anymore at the time of our study, no further information could be found.

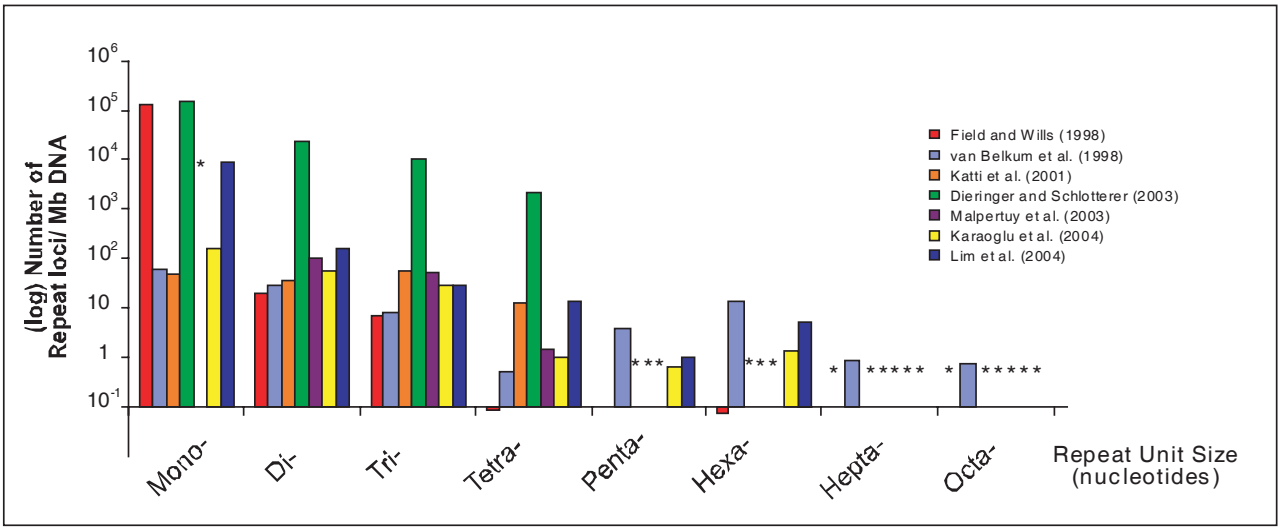


Figure 1. Microsatellite distribution in *S. cerevisiae*. Histogram shows the number of repeat loci per size class reported by each study. For details on parameter settings see Supplementary Table 1). *no data available.

same range of motifs, it is surprising to see such wide divergence in results. Here we discuss, that the crux of the problem derives from the different definitions of microsatellites used in each study. Differences in characteristics such as array length, unit size and purity inevitably transcribe into deviations in the parameter settings used in bioinformatic search tools, which subsequently lead to large discrepancies in results.

Minimum array length

Historically, the preferred size for microsatellites selected as genetic markers has been a minimum of five repeats (Selkoe and Toonen, 2006). However, the minimum array length required for strand slippage to occur is much lower. Rose and Falush, (1998) determined a critical length at around eight nucleotides based on microsatellite distribution in yeast, while Lai and Sun, (2003) approximated a minimum threshold of four copies for di-, tri-, tetra-, penta- and hexanucleotides and at least nine copies for mononucleotides for humans. In practice, however, the actual *in silico* detection of short repeats may be restricted by the minimum resolution of the search algorithm, e.g. 10 or 11 nucleotides in the case of Tandem Repeats Finder (Benson, 1999) used by Malpertuy et al. (2003). Within our meta-analysis the differences in minimum cut-off length explain most of the variance: studies applying a low length threshold, e.g. in the case of mononucleotides around 2–5bp (Dieringer and Schlotterer, 2003; Field and Wills, 1998; Lim, et al. 2004), harvest high repeat frequencies, whereas studies applying a higher threshold of 10 or 20bp report far fewer microsatellites (Karaoglu et al. 2005; Katti et al. 2001; van Belkum et al. 1998) (see Table 1).

Repeat unit size

Di-, tri- and tetranucleotide repeats dominate the literature because they have been found most frequently in the genome and are useful genetic markers (Jarne and Lagoda, 1996). Mononucleotides, whilst common, have been largely avoided as they cause problems during amplification (Selkoe and Toonen, 2006). However, from a mechanistic point of view, microsatellites are characterized by high levels of length polymorphism caused by DNA strand slippage, which can occur in repeat arrays composed of units that range from 1 to ~10bp in length (Armour et al. 1999; Jeffreys et al. 1994; Levinson and Gutman 1987b; Sia et al. 1997). Definitions of the motif length required to constitute a microsatellite vary in the literature: i.e. 1–6bp (Goldstein and Pollock, 1997), 1–5bp (Chambers and MacAvoy, 2000), 2–6bp (Schlotterer et al. 1998), or even 2–8bp (Armour et al. 1999). The same spread is reflected in our study survey: out of seven analyzed studies, one study excludes mononucleotide repeats (Malpertuy, Dujon, and Richard, 2003), only four studies report numbers for penta- and hexanucleotides, and only one examines hepta- and octanucleotides (van Belkum et al. 1998) (see Table 1 for search parameters).

Purity and internal structure of the array

So far, the majority of *in silico* searches have investigated only perfect microsatellites as they are computationally easier to detect. However, perfect microsatellites are not the only type of microsatellites. In fact, a repeat array might be classified as perfect (identical copies), imperfect (mismatches and indels are allowed) or compound/complex (array includes different motifs) (Buschiazzo and Gemmell, 2006; Chambers and MacAvoy, 2000). For most of the recent repeat detection tools, the level of imperfection can be varied as a parameter within the search. Despite this, Katti et al. (2001) and Malpertuy et al. (2003) are the only studies in our survey that allowed imperfections: a mismatch every 10th nucleotide, and succeeding mismatches after the first five perfect copies, respectively. While the available data do not allow us to detect a correlation between more or less stringent search criteria and high or low reported microsatellite frequencies, it appears logical that the inclusion or exclusion of imperfections in search parameters will influence the results of genomic comparisons.

Computational approach and genome build

There are additional, more subtle variables in the search that are rooted within the bioinformatic approach itself. Peculiarities of the underlying algorithm, such as combinatorial treatment of repeats in the identification procedure and/or redundancy filtering of overlaps or internal repetitions, may profoundly affect the overall pattern reported. Within our dataset, four studies (Katti et al. 2001; Lim et al. 2004; Malpertuy et al. 2003; van Belkum et al. 1998) apply the same minimum length threshold of 20bp in the case of tetranucleotides, but report frequencies of 0.5, 1.5, 12.6 and 13 repeats/Mbp, respectively. Comparing the documentation for the search approaches (Table 1) suggests that studies using different algorithmic approaches report varying repeat frequencies. Unfortunately, details of parameter settings and the structure of the applied algorithm are not consistently published, thereby precluding detailed comparisons.

Different sequence builds and the inclusion of the mitochondrial genome (mtDNA) in the sequence analyzed can also contribute to variation in results. We ran TRF in default mode on three different *S. cerevisiae* genome builds and found no significant variation in the total numbers, types and distributions of the microsatellites reported (Supplement 1). However, a significantly higher frequency of microsatellites was detected within the mitochondrial genome compared to the nuclear genome (Supplement 2) and the inclusion or exclusion of this genome in comparisons would result in a modest difference between studies.

Conclusion

The issue of how to exactly define a microsatellite is a long argued subject, upon which researchers have not yet reached consensus. Differences in parameters used in repeat detection, especially minimum array length, lead to large systematic biases in study results, where variations in microsatellite frequency can reach the extent of several magnitudes among studies even within the same genome.

Several authors have put forward microsatellite definitions, varying mainly based on their research background. First, describing types of repeats with respect to the degradation and complexity of the array subdivisions can be quite specific, such as in forensic and medicine (Urquhart et al. 1994), focusing on mutational behaviors of individual loci and alleles. We are predominately concerned with genomic analysis and propose therefore only three types of microsatellite spanning mono-hexanucleotides: perfect (repeat copies 100% identical), imperfect (mismatches and indels incorporated) and complex/compound (consist of several motifs, potentially with mismatches). Second, minimum array length has been traditionally defined by the occurrence of strand slippage events and the extent of the resulting microsatellite polymorphism. This has led to analyses employing either stacked thresholds that depend on repeat size (for example see Table 1) or length classes, e.g. microsatellites class I: 12 $<$ 20nt, microsatellite class II: $>$20nt (Temnykh et al. 2001).We suggest the following thresholds to start with, after Lai and Sun (2003): 12nt for mono-trinucleotides, 16nt for tetranucleotides, 20nt for pentanucleotids and 24nt for hexanucleotides. Absolute minimum thresholds for slippage events, tend to be group specific (between 8–15nt) and need to be adjusted individually for each species to eliminate background noise, i.e. random occurrences of microsatellites, from true over- or under representation.

Ideally, future studies ensure that all data are gathered and analyzed in a consistent manner, which should enable a consensus approach to emerge within the literature. However, due to the potential intricacies of microsatellite distribution in different genomic architectures, this might not always be possible in an absolute manner. Therefore, we encourage all authors to report their parameter settings and algorithms in detail (including the underlying reasoning), to enable sensible comparisons across studies. The importance of the issue can not be emphasized enough in the genomic era, where cross-species comparisons are the tools of trade.

Abbreviations

nt: nucleotide; kb: kilo base.

Acknowledgements

This work was supported by the Royal Society of New Zealand MARSDEN Fund UOC-202.

References

Abajian, C. 1994. *Sputnik.* http://espressosoftware.com/pages/sputnik.jsp

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L. and Batzer, M.A. 1995. Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, 29:136–44.

Armour, J.A.L, Alegre, S.A., Miles, S., Williams, L.J. and Badge, R.M. 1999. Minisatellites and mutation processes in tandemly repetitive DNA. In Goldstein, D. and Schlötterer, C. (eds), *Microsatellites: Evolution and Applications*, Oxford University Press, New York, 24–33.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27:573–80.

Buschiazzo, E. and Gemmell, N.J. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, 28:1040–50.

Chambers, G.K. and MacAvoy, E.S. 2000. Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology B.-Biochemistry and Molecular Biology*, 126:455–76.

Dieringer, D. and Schlotterer, C. 2003. Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. *Genome Research*, 13:2242–51.

Field, D. and Wills, C. 1998. Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S-cerevisiae, result from strong mutation pressures and a variety of selective forces. *Proceedings of the National Academy of Sciences of the United States of America*, 95:1647–52.

Goldstein, D.B. and Pollock, D.D. 1997. Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. *Journal of Heredity*, 88:335–42.

Goldstein, D.B. and Schlötterer, C. 1999. *Microsatellites : Evolution and Applications*, Oxford University Press, Oxford; New York.

Jarne, P. and Lagoda, P.J.L. 1996. Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, 11:424–9.

Jeffreys, A.J., Tamaki, K., Macleod, A., Monckton, D.G., Neil, D.L. and Armour, J.A.L. 1994. Complex Gene Conversion Events in Germline Mutation at Human Minisatellites. *Nature Genetics*, 6:136–45.

Karaoglu, H., Lee, C.M.Y and Meyer, W. 2005. Survey of simple sequence repeats in completed fungal genomes. *Molecular Biology and Evolution*, 22:639–49.

Kashi, Y. and King, D.G. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, 22:253–59.

Katti, M.V., Ranjekar, P.K. and Gupta, V.S. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution*, 18:1161–67.

Lai, Y.L. and Sun, F.Z. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution*, 20:2123–31.

Levinson, G. and Gutman, G.A. 1987b. Slipped-Strand Mispairing—a Major Mechanism for DNA-Sequence Evolution. *Molecular Biology and Evolution*, 4:203–21.

Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. 2004. Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution*, 21:991–1007.

Lim, S., Notley-McRobb, L., Lim, M. and Carter, D.A. 2004. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genetics and Biology*, 41:1025–36.

Malpertuy, A., Dujon, B. and Richard, G.F. 2003. Analysis of microsatellites in 13 hemiascomycetous yeast species: Mechanisms involved in genome dynamics. *Journal of Molecular Evolution*, 56:730–41.

Rose, O. and Falush, D. 1998. A threshold size for microsatellite expansion. *Molecular Biology and Evolution*, 15:613–5.

Schlotterer, C., Ritter, R., Harr, B. and Brem, G. 1998. High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution*, 15:1269–74.

Selkoe, K.A. and Toonen, R.J. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, 9:615–29.

Sia, E.A., JinksRobertson, S. and Petes, T.D. 1997. Genetic control of microsatellite stability. *Mutation Research-DNA Repair*, 383:61–70.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and McCouch, S. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research*, 11:1441–52.

Toth, G., Gaspari, Z. and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research*, 10:967–81.

Urquhart, A., Kimpton, C.P., Downes, T.J. and Gill, P. 1994. Variation in Short Tandem Repeat sequences a survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine*, 107:13–20.

van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H. 1998. Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and Molecular Biology Reviews*, 62:275–93.

Detecting Microsatellites in Genome Data: Variance in Definitions and Bioinformatic Approaches Cause Systematic Bias

Angelika Merkel and Neil J. Gemmell

Supplementary Material

Table S1. Variation in TRF results* between genome builts

Date genome built	1/01/1998	1/10/2003	30/11/2006
Total sequence size (nuclear), nt	12069303	12070521	12070899
Repeats found with TRF (default)	406	407	406

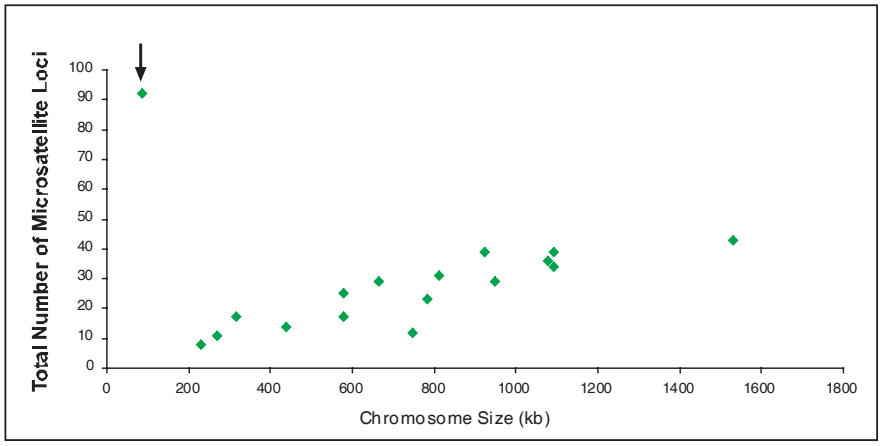


Figure S1. Varition in microsatellite abundance between different chromosome and mtDNA (↓). Note the roughly linear relationship between loci number and chromosome size with mtDNA (↓) as outlier.
Sequences were downloaded from ftp at SGD (ftp://genome-ftp.stanford.edu/pub/yeast/sequence/NCBI_genome_source).
*TRF default parameters: 2 7 7 80 10 50 6 (minimum length: 25nt)

Briefings in Bioinformatics Advance Access published July 10, 2008

doi:10.1093/bib/bbn028

Detecting short tandem repeats from genome data: opening the software black box

Angelika Merkel and Neil Gemmell

Submitted: 21st March 2008; Received (in revised form): 6th June 2008

Abstract

Short tandem repeats, specifically microsatellites, are widely used genetic markers, associated with human genetic diseases, and play an important role in various regulatory mechanisms and evolution. Despite their importance, much is yet unknown about their mutational dynamics. The increasing availability of genome data has led to several *in silico* studies of microsatellite evolution which have produced a vast range of algorithms and software for tandem repeat detection. Documentation of these tools is often sparse, or provided in a format that is impenetrable to most biologists without informatics background. This article introduces the major concepts behind repeat detecting software essential for informed tool selection. We reflect on issues such as parameter settings and program bias, as well as redundancy filtering and efficiency using examples from the currently available range of programs, to provide an integrated comparison and practical guide to microsatellite detecting programs.

Keywords: *microsatellite; tandem repeat; genome; algorithm; software; method; comparison*

INTRODUCTION

Microsatellites are short tandemly repeated (STR) DNA sequences of 1–6 bp unit length. Ubiquitously distributed in eukaryotic and prokaryotic genomes and highly polymorphic they rapidly became the current genetic marker of choice. Their usage is wide and includes genetic mapping, population genetic analysis, DNA forensics and phylogenetics [1]. More recently, microsatellite mutational dynamics have gained increasing interest as they have been shown to play a role in human genetic disorders [2] and may have significant roles in the regulation of gene expression [3, 4]. For example, microsatellites have been found to be major effectors of morphological evolution in dogs and distinctive social behaviour in voles [5, 6].

With the sequencing of the first eukaryotic genome in 1996, the yeast *Saccharomyces cerevisiae* [7], a new *in silico* approach based on bioinformatic tools opened up for studying microsatellite evolution. Now, microsatellites could easily be detected from genomic data instead of using the cost- and labour-intensive laboratory approaches involving probe hybridization. To date, numerous algorithms and related software have been developed to explore microsatellite distribution in prokaryotes and eukaryotes, with investigations ranging from studies of regional distribution bias to putative association with genomic features [8–14]. These days, most sequence analysis packages or genome browsers incorporate by default some form of tandem repeat finder, e.g. equicktandem and etandem at EMBOSS, repeat in the GCG-package and TandemRepeatFinder (TRF) at the University of California at Santa Cruz (UCSC) [15]. Likewise so called repeat masking and low complexity filtering tools, such as RepeatMasker [16] or DUST/SIMPLE [17, 18], are now standard components of sequence similarity search tools, like

Corresponding author. Angelika Merkel, School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8041, New Zealand. Tel: +64 (0)3 364 2987 ext 7048; Fax: +64 (0) 3 364 2509; E-mail: ame52@student.canterbury.ac.nz

Angelika Merkel is in the final year of her PhD thesis studying microsatellite evolution in the yeast genome. She pursued her undergraduate studies at the University of Constance (Germany), and has been a postgraduate student at the University of Canterbury (New Zealand) since 2004.

Neil Gemmell is Professor of Reproduction and Genomics at the University of Otago. Currently, a major facet of research in his lab is focused on the use of comparative genomic approaches to investigate the evolution of microsatellites, a question that is of paramount importance to the myriad of researchers using these molecular markers.

BLAST and BLAST-like applications, to reduce redundancy and speed up genome-wide pattern match searches. Finally, several repeat specific databases have been established to serve as references for such diverging objectives as studying model organisms, e.g. TandemRepeatDatabase [53], and EuMicrosatdb [19], and DNA forensics, e.g. STRbase [20]. There are also numerous programs that detect repeats in protein sequences, some of which share feature with DNA-orientated detection algorithms [21, 22].

Two recent studies further denote the popularity of these tools. Leclercq *et al*. [23] show a bias in repeat detection between algorithms, comparing some of the most commonly used tandem repeat finding programs, and Sharma *et al*. [24] give a first overview over the available software for microsatellite detection while illustrating facets of microsatellite distribution in eukaryotic genomes. Nevertheless, for most biologists the variety of software tools is rather overwhelming and selecting an application appropriate for the question posed becomes a challenge. Here we describe the fundamental concepts implemented in STR finding algorithms in order to provide a first practical guide to these commonly applied tools. We use examples from currently available software and discuss the utility of various applications for specific purposes. We see this information as an important step in moving biologists to develop selective approaches for microsatellite and repeat sequence detection, rather than the more common implementation of software as a mysterious black box.

SEARCH ALGORITHMS

In simple terms, a repeat finder program consists of three components: a detection unit, a filter component and the output compartment (Figure 1). The detection unit, harboring the search algorithm, is the core determinant of the overall time and space efficiency of the program. Based on certain selection criteria (statistics, scoring matrix) it detects patterns (motifs, repeats) specified under the users' input parameters. The resulting candidate repeats then undergo a filtering step to eliminate various types of redundancy. Outputs and utilities can vary widely between programs, i.e. including detailed information on the individual repeat, summary statistics or even additional modules for subsequent analysis (primer design, clustering or alignment).

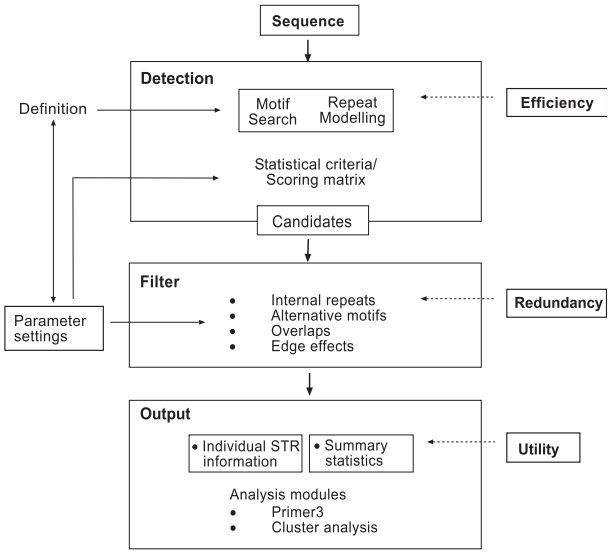


Figure 1: Schematic overview of a generic repeat finder algorithm.

Approaches

From a user's point of view, the identification of tandem repeats within a larger sequence takes two maxims: First, whether the search is going to be pattern specific, or unspecific (based on the repetitive nature of a sequence only); and second, what type of repeats will be searched for (perfect, imperfect or complex repeats) (see Table 1 for examples).

From a programmer's point of view, the most straight forward approach to identify repeats is to search for specified sequences or motifs. In principle this can be achieved using any text editor, but practically, most searches investigate at least a set of motifs in very large texts, i.e. whole genomes. For some applications, like TROLL [25], an application based on the Aho-Corasick algorithm [25], the user can provide a list of motifs in a separate input file which are then searched against the query all at once. Similarly, but based on a local alignment strategy, RepeatMasker [16] uses a list of pre-selected common motifs, stored in a reference database called RepBase [26], to scan a query for these sequences. Here the reference pattern is aligned along a genomic sequence implementing a scoring matrix. If a match is encountered, the adjacent sequences are aligned and subsequently masked if they exceed a certain threshold. Both programs are effective in detecting a defined set of patterns in a sequence and are highly suitable for selective motif searches, but these are not effective substitutes for more comprehensive search tools (e.g. TRF [15] or Sputnik [27],

Table 1: Commonly employed terms for microsatellites/STRs

Biological definition	Mathematical/Computational description	Features	Example
Perfect	Exact match	100% identical copies	$(A)_n$, $(ATC)_n$
Imperfect	Approximate-hamming Distance (HD)	substitutions (= mismatches)	$(AC)_nAT(AC)_m$
Interrupted*	Approximate-edit distance (ED)	substitutions, insertions, deletions (= interruptions)	$(ACG)_nT(ACG)_m$, $(AT)_nCGAG(AT)_m$
Compound/complex	'Fuzzy'	multiple motifs, periods, substitutions	$(ACG)_nT(TC)_m$, ATcgcıATggcıATtccıATcgg

*Interrupted repeats are often included in imperfect repeats.

see below) that can be used for example to estimate genome-wide repeat content.

Regular expressions are an efficient and hence popular way to search for repeats of a certain size and a large number of patterns. A regular expression describes a set of strings or patterns according to certain rules, such as the incorporation of wildcards into the motif at a fixed frequency. A variety of software languages accommodate regular expressions in their syntax, but due to its powerful inbuilt regular expression search engine *regex* and its text processing capabilities, many repeat detecting algorithms that have been written in Perl, such as MsatFinder [28], SSRIT [29] and MISA [30]. Msatfinder even employs regular expression searches at various levels of speed and accuracy: (i) fast regular expression (sequence is searched only once) and (ii) regular expression (sequence is searched several times); the first variant being a faster but less precise search and the second variant being slightly slower but more accurate in detection.

The first combinatorial approach to identify microsatellites/STRs based only on repeat size, was implemented in the program Sputnik in 1994 [27]. Sputnik employs a recursive algorithm using sliding windows to detect repeats of 1–5 bp length by scanning through the sequence one base at a time, and checking subsequent bases for repeats. Matches of adjacent windows are evaluated by a scoring matrix. Initial repeats are extended and reported as long as they meet the minimum threshold. Poly [31], uses a similar base-by-base search, but differs from Sputnik [27] by searching for all window sizes at once instead of only searching for one pattern size at a time. The algorithm constructs accretive windows at each base of the input sequence, starting with the minimum pattern size. If there is no exact match to the preceding window, the window size is increased. Alternatively, if the maximum pattern size is reached and no match is detected, the starting position of the window shifts to the next base. However, both programs do not appear to differ remarkably in their execution times. Since its initial release, Sputnik has been modified several times to improve either search capacity or output flexibility [13, 32]. The latest development from the Sputnik family tree is SciRoKo [33], an extremely flexible tool, that incorporates fixed mismatch penalties as well as variable penalties (i.e. motif length⋆X).

Most of the approaches outlined above only search for very short tandem repeat such as microsatellites and/or employ very simple substitution models, if a substitution model is employed at all. However, as a consequence of the recognition of tandem repeats as an essential component of all genomes analyzed so far and the general observation that imperfect/complex repeats are more prevalent than perfect repeats, a large number of algorithms have been developed that model tandem repeats by employing the distance criteria (i.e. repeat size) as part of the search matrix itself. Such tools allow users to search for repeat sizes larger than microsatellites (e.g. minisatellites, 10 bp to ~100 bp) and to search for specific types or patterns of repetition (Table 1).

Amongst these, TRF [15] is probably the most common and widely used tool for finding tandem repeats and has provided the basis for many other such tools [34, 35]. Initially, the algorithm uses sliding windows to search for matching nucleotides separated by a common distance. Like the Smith–Waterman algorithm [36] it requires only partial matches between copies, called *k-tuple* matches (seeds). For each *k-tuple* match, the distance information and location are stored in an index. To select relevant candidates from the list a variety of statistical criteria are applied, which themselves are derived

from several probability distributions (pattern length; matching probability P_m, indel probability P_i and tuple size k). The result is not an exhaustive search but a sufficient one that in a heuristic manner enables reasonable fast processing of very large datasets, such as mammalian genomes.

ATR-hunter by Wexler *et al.* [35] takes a similar heuristic/statistical approach. In addition to indexing the distance and location of potential repeat copies, it utilizes a quality vector to describe the type of repetition. Applying scorings for matches and gaps of individual segments it is possible to find approximate repeats based on different similarity measures. Whereas TRF uses an alignment of each repeat copy to a consensus sequence as similarity measurement, ATR-hunter scores mutations between neighbouring copies or alternatively, the average similarity between all copies of the array, making it more flexible in detecting various types of repeats (Table 1).

Other applications have extended the concept of imperfect tandem repeats even further. TandemSwan [34] detects so called 'fuzzy' repeats, i.e. repeats that can differ in number of mismatches per copy, period and number of copies. Based on an autocorrelation analysis, adjacent windows are compared to each other. Each letter comparison of a neighbouring window receives a score and repeats are eventually identified via a minimum function. The actual output candidates are selected via *P*-value thresholds based on the level of divergence between copies and motif similarity. Similarly, Mreps [37] detects repeats composed of different motifs but is based on a seed extension technique instead. Here, initially exact repeats are detected which are then, dependent on a resolution parameter set by the user, maximal extended. All discovered hits undergo extensive redundancy treatment (see below) and are statistically verified based on a real distribution in a random DNA sequence.

Redundancy

Increasing the complexity and sensitivity of repeat detection is usually paralleled by increased redundancy in the discovered repeats, and thus the complexity of the analysis filter generally increases with the complexity of the search engine. Filtering is crucial for removing redundant output and particularly vital for accurate counts. However, the necessity for repeat filtering, and more importantly the type of filtering, should be determined based on the biological significance and research focus.

Fore instance, duplicated motifs such as $(ATAT)_2$ instead of $(AT)_4$, and permutations of the motif via alternative reading frames, such as AT versus TA, appear of no biological difference and can easily be discarded as redundant. Whether AT or TA will be reported as a motif is subject to the neighbouring mismatches in the sequence and the threshold settings of the search tool. Generally, such location dependent redundancy filtering is achieved within the algorithm through a list or buffer where all repeat positions are recorded and from which eventually only a single hit per position is reported. Nevertheless, motif identification can be troublesome in the case of imperfect or very degenerate repeats. TRF [15], for example, reports up to three possible motifs per locus allowing the user to manually check weather a motif has been correctly assigned to a repeat by the software, or not. This is potentially very useful when studying a particular motif type, but presents a major barrier to precise repeat counts and density estimates. Additional external redundancy filters may have to be applied if accurate counts are to be obtained (e.g. for genome-wide microsatellite coverage). Alternatively, such as in Sputnik [27] and SciRoko [33], permutations of a motif and the corresponding complementary motifs are grouped together in a natural sense [38]. The grouping of these motifs and their complementary motifs together has to be taken with caution if the research focus is on investigating microsatellite evolution, as some studies have shown strand preferences for certain motifs [13, 39]. Finally, merging of overlapping or adjacent repeats is yet another filter strategy, which is directed at certain repeat definitions, particularly compound or interrupted repeats, respectively. In some applications merging is optional (e.g. SciRoko [27], MsatFinder [28]), but in others, such as Mreps [37] merging is an integral component of the program and constitutes an additional purification step after a relaxed search. Here again, precise frequency estimates are traded-off for accurate motif distributions, and the choice of filter (or program) has to be made with respect to study purpose.

For example, if one was interested in the distribution pattern of $(AC)_n$ across various genomes, its frequency could be underestimated by merging or grouping. Programs like Star [40], TROLL [25] and IMEx [41] (pattern search optional) can eliminate

inferences from other motifs through a motif specific search. Alternatively, the same information could be retrieved via summary statistics (see below), provided merging and grouping options can be modified in the filter settings (Msatfinder [28], SciRoko [33]). On the other hand, if one was interested in overall microsatellite frequencies, such as occurrences per megabase (loci) or genome-wide coverage (nt), the merging of overlapping repeats is crucial while sorting of motifs becomes irrelevant.

Study bias – algorithms and parameter settings

Naturally, different approaches are likely to diverge slightly in their outcomes, and tandem repeat detecting software is no exception. Nevertheless, we recently conducted a meta-analysis on published microsatellite distribution in yeast [42] that showed a divergence of up to three orders of magnitude in the frequency of microsatellite motifs reported among seven studies. We showed that the observed discrepancies are predominantly due to different parameter settings between studies which themselves emerge from different definitions applied for microsatellites (e.g. minimum array length/repeat number, motif length, perfection/degeneration of the array). We further found a bias depending on the algorithm employed (Figure 2) mainly in number of repeats detected, size classes identified and length distribution. Complimentary findings have been reported by Leclercq *et al*. [23]. Here, the authors tested five repeat finding programs, namely TRF [15], Sputnik [27], Mreps [37], STAR [40] and RepeatMasker [16], across several eukaryotic genomes and found major divergence in the repeats detected depending on the program, and more significantly the parameter settings selected. For example the study shows, that, at extreme values Sputnik [27] detects an 80-fold amount of perfect

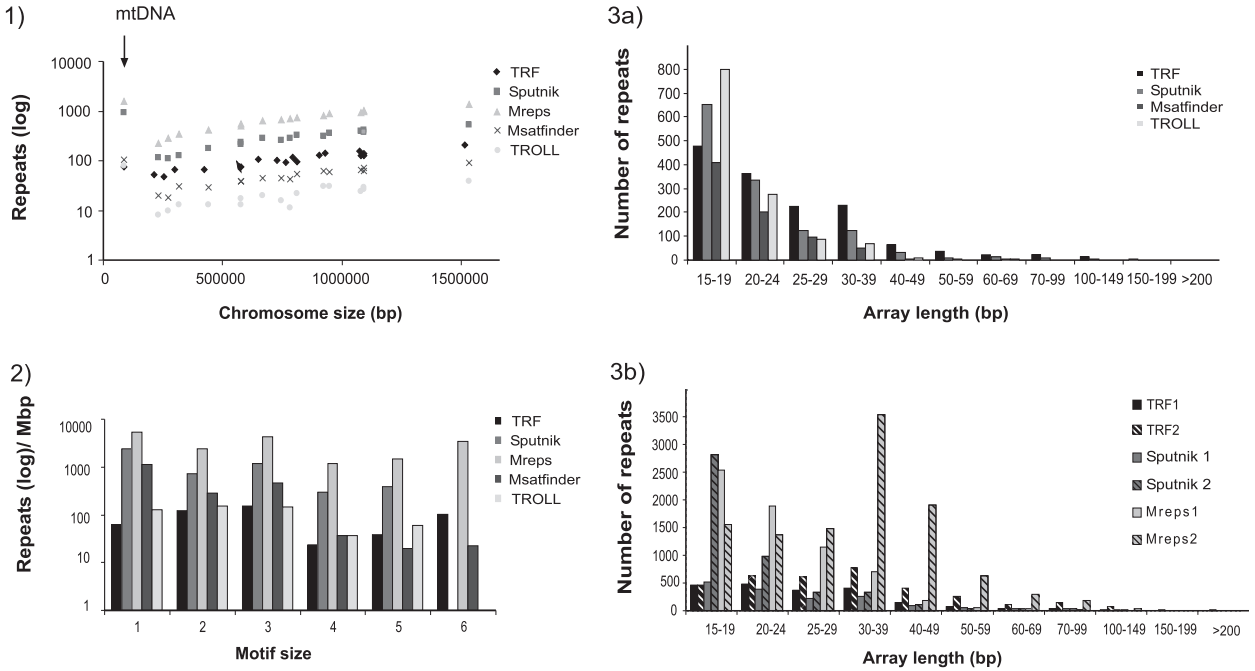


Figure 2: Comparison of microsatellite finding programs using the yeast genome. **1**) Distribution of microsatellites across the mitochondrial genome and all chromosomes. Note the overabundance of repeats in mtDNA compared to nuclear DNA. **2**) Distribution of repeat unit sizes. **3**) Distribution of array length under **a**) stringent parameter settings which detects mostly perfect repeats and **b**) relaxed parameter settings that allows for more imperfections in the repeat sequence. Noteworthy is the increased level of detection of intermediate length arrays for Mreps2, caused by Mreps inbuilt merging procedure. For **1**) and **2**) default parameters were used. For 3a) the parameter settings were: TRF: m = 2, mism = 7, indels = 7, pm = 80, pi = 10, score = 30; Sputnik: m = 1, mism = − 6, score = 10; Mreps: res = 3, TROLL = default, MsatFinder = default. Parameter settings for 3b) are: TRF1: m = 2, mism = 5, indels = 5, pm = 80, pi = 10, score = 30; Sputnik1: m = 1, mism = -3, score = 12; Mreps1: res = 3; TRF2: m = 2, mism = 3, indels = 5, pm = 80, pi = 10, score = 30; Sputnik2: m = 1, mism = −3, score = 5; Mreps2: res = 1. Minimum array length (filtered) for all searches = 15 nt, repeat size = 1–5 bp.

repeats detected by RepeatMasker on human chromosome X, and TRF [15] shows an 61% increase in detections between two different alignment weights (2,7,7 and 2,3,5). Nevertheless, the observed biases were consistent across different genomes; hence, it seems there is no sequence specific program bias.

At a glance, such reports seem alarming and fundamentally question the accuracy of *in silico* microsatellite detection. Nevertheless, the underlying mechanics of the discrepancies can be traced. Considering algorithms implementing a scoring matrix for repetitive sequence identification, the standard parameters are minimum array length, minimum score and alignment weights. Minimum length is the most critical parameter for repeat detection, because short microsatellites are highly overrepresented in the genome. Hence, detections increase exponentially with decreasing minimum length. Threshold scores determine mean length and number of repeats detected but also influence the average degree of perfection within repeats, as imperfections lower the score [23]. High threshold scores produce shorter and more perfect microsatellites, while lower threshold scores produce overall more, but on average longer and more imperfect repeats. In contrast, alignment weights (matches, mismatches, indels) predominantly extend or shorten already existing repeats, but only slightly increase the number of detections [23]. Finally, threshold scores and alignment weights modulate the detected frequencies for different repeat size in quite a complex fashion, due to different size classes exhibiting unequal degrees of imperfection (Figure 3).

The individual search engine employed may also have an effect on the type of repeat detected with regards to average length and/or the level of divergence in motif. TRF detects on average longer, but more imperfect repeats, whereas Sputnik detects shorter, but more perfect repeats (based on similar parameter settings and uniform

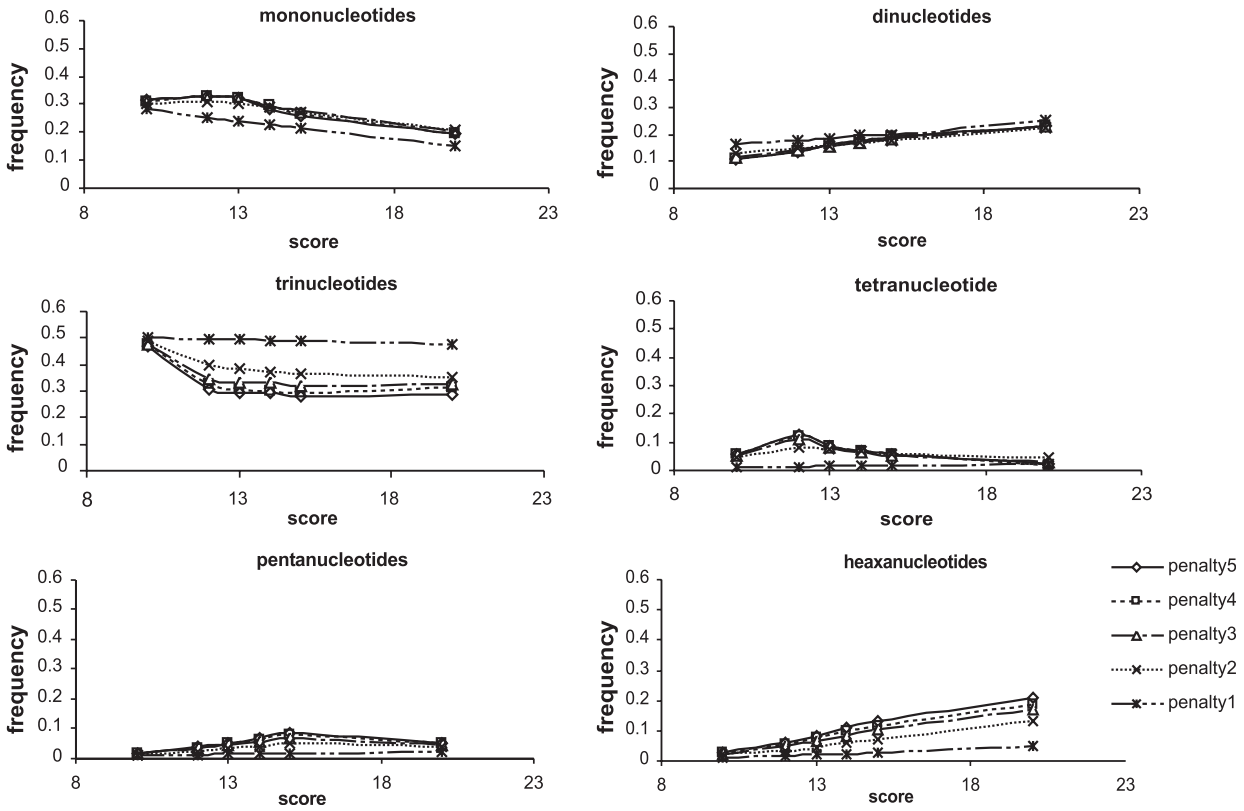


Figure 3: Influence of mismatch penalty and threshold score on different repeat sizes. High threshold scores and mismatch penalties increase di- and hexanucleotide frequency, as well as decreasing mononucleotide frequency. Tri- and tetranucleotide show temporary frequency changes at low scores, pentanucleotids at intermediate scores. Analysis was performed with SciRoKo [26] on the nuclear *S.cerevisiae* genome only (score = minimum score, penalty = fixed mismatch penalty, minimum array length = 8 nt or 3 repeats).

divergence estimates) (Figure 2). This difference among the programs is likely due to TRF creating the repeat alignment based on a consensus sequence whereas Sputnik compares neighbouring copies to each other. Mreps [37], which does not imply any minimum criteria for repeat identification such as score or length but a fixed seed size instead, shows no such bias, and detects repeats of equivalent degeneration regardless of their length (Figure 2). The longest and most divergent repeats are found by RepeatMasker [16] due to its pairwise alignment approach [23]. Finally, repeat finders for only perfect repeats, like Msatfinder [28] and TROLL [25], identify naturally shorter repeats than other programs. Overall, the positions of most repeats overlap between programs in similar proportion as numbers of overall repeats detected increase. Still, some repeats are unique and combining search approaches can yield higher sensitivity.

Practically, problems are commonly encountered when searching for very short repeats. Those can only be detected at very low thresholds when using certain programs. But low thresholds create usually a lot of background noise, made up of highly degraded 'microsatellites' that appear to be closer to random sequences than microsatellites of biological significance. One way this can be avoided is using an additional program that has a higher fidelity for shorter repeats, and subsequently combing both results. However, as a backdrop, an additional filtering step becomes necessary to eliminate overlapping repeats. To avoid methodological biases and verify their results a few studies have employed multiple searches using a variety of parameter settings, different algorithms, or both [43–45].

Efficiency

The issue of search efficiency becomes rapidly apparent when processing large datasets, such as whole genome data, on standard desktop machines or even laptops. The time and space requirements of a tandem repeat search algorithm are directly correlated with the intricacy of the search [46]. So algorithms for detecting exact repeats have the shortest running times exhibiting a linear time progression followed by algorithms detecting approximate repeats under the Hamming distance model (logarithmic running times). The most computationally costly algorithms are those that detect approximate repeats under the edit distance model (quadratic running times). Many string matching algorithms use dynamic programming routinely as a technique to increase processing speed.

On a structural level, the number of computations can be efficiently reduced by pre-processing of either the input sequence or, in the case of a motif search, the pattern itself. For example, TROLL [25] constructs in a pre-processing stage a keyword tree from the motif input file, which then can be used to search multiple sequences. A common technique for increased search speed is to transform the queried sequence into a complex data structure to enable fast look-ups. REPuter by Kurtz and Schleiermacher [47] incorporates suffix trees to search not only for tandem repeats but also large interspersed repeats. Far more exotic, STAR by Delgrande and Rivals [40] utilizes methods from the field of data compression to simplify the queried sequence. The sequence, together with the recording of mutations/alterations,

Table 2: Repeats, parameters and potential resources related to studies focusing on microsatellites and STRs

Study goal	Type of repeats searched for	Parameter settings applied	Suggested resources*
Amplification of microsatellites (Primer design), identifying polymorphic microsatellites (prediction or *in silico* allele scoring)	Polymorphic, i.e. long and perfect arrays	Stringent, high penalties, minimum score and thresholds	SSRprimer, IMex, MsatFinder, Misa, TRDB, VNTRfinder, PolyPredictR
Characterize genomic microsatellite distribution, study microsatellite evolution	All types, specific motifs/repeat unit sizes/array lengths	Various (study specific)	SciRoko, Sputnik, IMex, TRF, Misa, MsatFinder
Estimating genomic microsatellite content	All types, non- redundant loci	Relaxed, low penalties, minimum score and thresholds	SciRoko, Sputnik, MsatFinder
Eliminate/mask redundancy prone regions	All types	Highly relaxed	RepeatMasker, DUST, SIMPLE
Find STRs/VNTRs other than microsatellites, e.g. minisatellites	All types	Various (study specific)	TRF, Mreps, STRING, STAR, ATRhunter, TandemSwan, etandem, repeat

*N.B. this list of resources is not exhaustive.

Table 3: Search tools used for STR detection, overview of features and properties

Program	Script	Operating system	URL	User interface	Type(s) of repeat	Summary statistic/ locus Info	Flanking sequences	Pri-mer3	Reference
RepeatMasker	Perl	Unix	http://www.repeatmasker.org/	console/web	perfect, imperfect (RepBase)	no	no	no	[16]
Msatfinder	Perl	Linux	http://www.bioinf.ceh.ac.uk/msatfinder/	console/web	1–6 bp perfect (imperfect, compound optional)	yes	yes	yes	[28]
Misa	Perl	Linux	http://pgrc.ipk-gatersleben.de/misa/	console	1–6 bp perfect (interrupted optional)	yes	yes	yes	[30]
SSRIT	Perl	Linux	http://www.gramene.org/db/searches/ssrtool	web	2–10 bp perfect	no	no	no	[29]
Sputnik	C	Windows, Linux	http://espressosoftware.com/pages/sputnik.jsp	console	1–5 bp perfect, imperfect (HD)	no	yes	no	[27]
Sputnik I	C	Windows, Linux	http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/	console	1–5 bp perfect, imperfect (HD)	no	yes	no	[13]
Sputnik II	C	Windows, Linux	http://cbi.labri.u-bordeaux.fr/outils/Pise/sputnik.html	console/web	1–5 bp perfect, imperfect (HD)	no	yes	no	[32]
Poly	Phyton	Linux	http://bioinformatics.org/poly/	console	1–4 bp perfect	yes	no	no	[31]
TRF	?	Windows, Linux, Mac	http://tandem.bu.edu/trf/trf.html	gui/web	1–2000 bp imperfect (ED)	yes	no	no	[15]
ATRhunter	C	Windows, Sun/Solaris, Linux	http://bioinfo.cs.technion.ac.il/atrhunter/ATRHunter.htm	web	1–500 bp imperfect (ED)	no	no	no	[35]
TandemSWAN	C, C++	Windows, Linux	http://strand.imb.ac.ru/swan/	console/web	3–100 bp fuzzy repeats	no	no	no	[34]
Mreps	ANSI C	Windows, Linux, Mac	http://bioinfo.lifl.fr/mreps/	console/web	1–? bp fuzzy repeats	no	no	no	[37]
STAR	?	Linux, Sun and Mac, Windows	http://atgc.lirmm.fr/star/	web	1–9 bp imperfect (ED), motif specific	no	?	no	[40]
STRING	C	UNIX, Windows, Mac	http://www.caspur.it/~castri/STRING/index.htm	console	Imperfect (HD)	yes	?	no	[52]
TROLL	C++	Unix, Linux	http://finder.sourceforge.net	console	perfect (motif file)	no	no	no	[25]
IMEx	C	Linux	http://203.197.254.154/IMEX/index.html	console/web	1–6 bp perfect, imperfect (ED)	yes	yes	yes	[41]
SciRoko	C#	Linux, Unix, Solaris, Free BSD, Mac	http://www.kofler.or.at/bioinformatics/SciRoKo/index.html	gui/web	perfect, imperfect, compound	yes	yes	no	[33]
etandem/ equicktandem	?C	Unix, Windows, Mac	http://emboss.sourceforge.net/	gui/console	perfect, imperfect	no	no	no	EMBOSS
findpattern/repeat	?C	Unix, Windows, Mac	http://www.accelrys.com/products/gcg/index.html	gui	perfect, imperfect	no	no	no	GCG-package

(continued)

Table 3: Continued

Database	Algorithm	Genomic Data	URL	Features	Reference
Repbase	Reference database	Eukaryotes	http://www.girinst.org/repbase/update/index.html	Transposable elements, simple repeats, pseudogenes	[26]
TRDB	TRF	Eukaroytes (44)	http://tandem.bu.edu/cgi-bin/trdb/trdb.exe	Personal project space, flankings equences, primer design, cluster analysis, TF binding site prediction, graphical representaion	[53]
EuMicroSatdb	Misa	Eukaryotes (31)	http://veenuash.info/web/index.htm	Batch download, flanking sequence, compound microsatellites, genomic position (intron/exon, intergenic, upstream region)	[19]
InSatdb	TRF	Insects (5)	http://210.212.212.8/PHP/INSATDB/home.php	Batch download, flanking sequence, compound microsatellites,GC-content, genomic position(intron/exon,intergenic, upstream region)	[54]
ABCC GRID Database	STR finder	UCSC	http://grid.abcc.ncifcrf.gov.	Sequence, feature query, gff-format, view at UCSC	[55]
Trbase	TRF	Human	http://trbase.ex.ac.uk/	Search for tandem repeats by gene/deseaseassociation, genomic position	[56]
SSRPrimer & SSR Taxonomy Tree	Sputnik	GenBank	http://bioinformatics.pbcbasc.latrobe.edu.au/ssrdiscovery.html.	Primer design, taxonomy search, visualization	[49]
IMEx	IMEx	Prokaryotes, virus	http://203.197.254.154/IMEX/index.html	See IMEx	[41]
VNTRfinder & PolyPredictR	TRF	various	http://www.bioinformatics.rcsi.ie/vntrfinder/	Detection/scoring of homologous alleles is multiple sequences, polymorphism prediction, repeat type, flanking sequences, gene diversity/heterozygosity	[50]

Note: '?' = unknown, not indicated in the references.

is transformed into a significance distribution. Repeats are subsequently detected as maxima in the distribution. The authors claim that the method also has the advantage to allow pattern size independent scoring (see above).

Flexibility and utility

Parameter flexibility, output options and other utilities vary widely with the available software. As user knowledge, sophistication and needs increase, fixed or flexible parameters might be preferred. A number of programs offer besides the default settings a hierarchy of different search levels, such as basic, intermediate and advanced with increasing amounts of parameter flexibility (IMEx [41], ATRhunter [35], Msatfinder [28]).

With regards to the many fold output options and additional functions available, program selection at this point should be made with the prime focus on the downstream analysis requirements (Table 2). All programs report at a minimum genomic position and the type or sequence of the microsatellite. Most programs supply further information about the microsatellite such as length, size class, base count, flanking sequence, GC-content of flanking sequence, and, in the case of imperfect repeats, some measure of imperfection, i.e. matches, mismatches, indels, percentage perfection of or even an entropy indication of the sequence in TRF [15]. A few programs provide summary statistics, e.g. total count, base coverage/density, average length, size class and motif abundance and some software also contain additional applications like Primer3 [48], designing primers automatically from the flanking sequence or modules for cluster analysis (Table 3). Hence, if the primary goal is primer design an application like IMex [41], MsatFinder [28], SSRPrimer [49] or Misa [30], that includes a Primer3 module, is best suited to the task and depending on the amount of sequence data to be examined a stand-alone version might be chosen over the web-interface. Local stand-alone versions generally process large datasets much faster than web-based counter parts, whereas web-based versions spare the user the time- and resource-consuming software install, and are sufficient for a small number of queries. On the other hand, if the research focuses on microsatellite distribution, such as for the purpose of characterizing microsatellite abundance or exploring genome architecture, the use of a stand-alone version providing a range of summary statistics-detailed locus information and fully flexible parameter settings that is almost mandatory. SciRoko [33], TRF [15], Sputnik [27], and others (Table 3) are all good choices for such tasks. Some specialized applications, such as VNTRfinder and PolyPredictR, also allow the prediction of potential allele variations or directly evaluate these using either preset rules for polymorphism detection or a combination of TRF and sequence alignment methods (e-pcr or BLAST), respectively [44, 50, 51]. A last source of microsatellite data and analysis tools are the purpose built databases for repetitive sequences. Several large microsatellite databases have already been established by pre-screening whole genome sequences for repeats (Table 3) and some genome browsers display microsatellite data routinely as an individual feature track, e.g. tracks in the UCSC genome browser created by RepeatMasker and TRF (http://genome.ucsc.edu/).

CONCLUSION

Applications for detecting microsatellites and other STRs are many and diverse. Key structural differences exist among these in terms of search engines, filter and utilities. Program resolution varies, and a methodological bias is observed among programs that are especially pronounced when parameter settings vary. Caution has to be taken when choosing parameters if comparable results are to be obtained among studies. Microsatellite distribution in terms of frequency or coverage and over-/under-representation of certain characteristics, such as motifs, should be interpreted with respect to the approach, i.e. repeat type or definition, and candidate validation statistics/filter. Finally, users may choose an application based on the repeat type, i.e. the repeat characteristic investigated, the efficiency and utility of the program, such as parameter flexibility, implementation (gui/web) and modules available for additional analysis.

Key Points

- Programs for STR detection vary significantly in repeat definition, search algorithm and filtering method.
- A detection bias between algorithms and especially parameter setting is observed.
- Minimum repeat array length and overall purity thresholds, i.e. the number of mismatches and, or, indels allowed per array, are critical parameters for efficient and accurate microsatellite detection.
- The study purpose, e.g. marker development as opposed to characterization of microsatellite abundance, is a key determinant in terms of tool selection with respect to program flexibility and utilities required.

References

1. Goldstein DB, Schlötterer C. *Microsatellites: Evolution and Applications*. Oxford: Oxford University Press, 1999.
2. Pearson CE, Edamura KN, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* 2005;**6**: 729–42.
3. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* 2006;**22**: 253–9.
4. Moxon ER, Wills C. DNA microsatellites: agents of evolution? *Sci Am* 1999;**280**:94–9.
5. Fondon JW, Garner HR. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci USA* 2004;**101**:18058–63.
6. Hammock EAD, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 2005;**308**:1630–4.
7. Goffeau A, Barrell BG, Bussey H, *et al*. Life with 6000 genes. *Science* 1996;**274**:546–67.
8. Dieringer D, Schlotterer C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res* 2003;**13**:2242–51.
9. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol* 2001;**18**:1161–7.
10. Li YC, Korol AB, Fahima T, *et al*. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* 2002;**11**:2453–65.
11. Li YC, Korol AB, Fahima T, *et al*. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 2004;**21**:991–1007.
12. Lim S, Notley-McRobb L, Lim M, *et al*. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genet Biol* 2004;**41**:1025–36.
13. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 2002;**30**:194–200.
14. van Belkum A, Scherer S, van Alphen L, *et al*. Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* 1998;**62**:275–93.
15. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**:573–80.
16. Smit AFA, Green P. RepeatMasker. 1996. Available at: http://www.repeatmasker.org/.
17. Alba MM, Laskowski RA, Hancock JM. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* 2002;**18**:672–8.
18. Hancock JM, Armstrong JS. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* 1994;**10**: 67–70.
19. Aishwarya V, Grover A, Sharma PC. EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics* 2007;**8**:225.
20. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res* 2001;**29**:320–2.
21. Depledge DP, Dalby AR. COPASAAR–a database for proteomic analysis of single amino acid repeats. *BMC Bioinformatics* 2005;**6**:196.
22. Kalita M, Ramasamy G, Duraisamy S, *et al*. ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics* 2006;**7**:336.
23. Leclercq S, Rivals E, Jarne P. Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics* 2007;**8**:125.
24. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol* 2007;**25**:490–8.
25. Castelo AT, Martins W, Gao GR. TROLL-Tandem Repeat Occurrence Locator. *Bioinformatics* 2002; **18**:634–6.
26. Jurka J, Kapitonov VV, Pavlicek A, *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005;**110**:462–7.
27. Abajian C. Sputnik. 1994. Available at: http://espressosoftware.com/pages/sputnik.jsp.
28. Thurston MI, Field D. Msatfinder: Detection and Characterization of Microsatellites. Oxford: Centre for Ecology and Hydrology, 2005. Available at: http://www.genomics.ceh.ac.uk/msatfinder.
29. Temnykh S, DeClerck G, Lukashova A, *et al*. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* 2001;**11**:1441–52.
30. Thiel T, Michalek W, Varshney RK, *et al*. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet* 2003;**106**:411–22.
31. Bizzaro JW, Marx KA. Poly: a quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. *BMC Bioinformatics* 2003;**4**:22.
32. La Rota M, Kantety RV, Yu JK, *et al*. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 2005;**6**:23.
33. Kofler R, Schlotterer C, Lelley T. SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* 2007;**23**:1683–5.
34. Boeva V, Regnier M, Papatsenko D, *et al*. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* 2006;**22**:676–84.
35. Wexler Y, Yakhini Z, Kashi Y, *et al*. Finding approximate tandem repeats in genomic sequences. *J Comput Biol* 2005; **12**:928–42.
36. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
37. Kolpakov R, Bana G, Kucherov G. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* 2003;**31**:3672–8.
38. Jin L, Zhong YX, Chakraborty R. The exact numbers of possible microsatellite motifs. *Am J Hum Genet* 1994;**55**: 582–3.
39. Freudenreich CH, Stavenhagen JB, Zakian VA. Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Mol Cell Biol* 1997;**17**: 2090–8.

40. Delgrange O, Rivals E. STAR: an algorithm to search for tandem approximate repeats. *Bioinformatics* 2004;**20**: 2812–20.
41. Mudunuri SB, Nagarajaram HA. IMEx: imperfect microsatellite extractor. *Bioinformatics* 2007;**23**:1181–7.
42. Merkel A, Gemmell N. Detecting microsatellites in genome data: variance in definitions and bioinformatic approaches cause systematic bias. *Evol Bioinform* 2008;1–6.
43. Kelkar YD, Tyekucheva S, Chiaromonte F, *et al*. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* 2008;**18**:30–8.
44. Naslund K, Saetre P, von Salome J, *et al*. Genome-wide prediction of human VNTRs. *Genomics* 2005;**85**:24–35.
45. O'Dushlaine CT, Edwards RJ, Park SD, *et al*. Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol* 2005;**6**:R69.
46. Landau GM, Schmidt JP, Sokol D. An algorithm for approximate tandem repeats. *J Comput Biol* 2001;**8**:1–18.
47. Kurtz S, Choudhuri JV, Ohlebusch E, *et al*. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* 2001;**29**:4633–42.
48. Rozen S, Skaletsky H. Primer3 on the WWW for general user and for biologists programmers. In: Krawetz S, Misener S (eds). *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Totowa, NJ: Humana Press, 2000, 365–86.
49. Jewell E, Robinson A, Savage D, *et al*. SSRPrimer and SSR taxonomy tree: biome SSR discovery. *Nucleic Acids Res* 2006;**34**:W656–9.
50. O'Dushlaine CT, Shields DC. Tools for the identification of variable and potentially variable tandem repeats. *BMC Genomics* 2006;**7**:290.
51. Wren JD, Forgacs E, Fondon JW, *et al*. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* 2000;**67**:345–56.
52. Parisi V, De Fonzo V, Aluffi-Pentini F. STRING: finding tandem repeats in DNA sequences. *Bioinformatics* 2003;**19**: 1733–8.
53. Gelfand Y, Rodriguez A, Benson G. TRDB–the tandem repeats database. *Nucleic Acids Res* 2007;**35**:D80–7.
54. Archak S, Meduri E, Kumar PS, *et al*. InSatDb: a microsatellite database of fully sequenced insect genomes. *Nucleic Acids Res* 2007;**35**:D36–9.
55. Collins JR, Stephens RM, Gold B, *et al*. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics* 2003;**82**:10–19.
56. Boby T, Patch AM, Aves SJ. TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics* 2005;**21**:811–16.

Evolution of Microsatellite DNA

Iris M Vargas Jentzsch, *University of Canterbury, Christchurch, New Zealand*
Andrew Bagshaw, *University of Canterbury, Christchurch, New Zealand*
Emmanuel Buschiazzo, *University of Canterbury, Christchurch, New Zealand*
Angelika Merkel, *University of Canterbury, Christchurch, New Zealand*
Neil J Gemmell, *University of Canterbury, Christchurch, New Zealand*

Advanced article

Article Contents

Online posting date: 30th April 2008

Microsatellites are highly mutable tandemly repeated sequences that are ubiquitously distributed in bacterial and eukaryotic genomes. Microsatellites became the preferred molecular marker for a variety of applications under the basic assumption that they are selectively neutral. However, the simplicity of this assumption contrasts with the observed variability of mutation rates across microsatellite loci and with the increasing evidence supporting microsatellite functionality. The evolutionary importance of microsatellites is only recently being uncovered with the intense study of their impact on the regulation of gene expression and the interaction among genomic structures.

Introduction

Microsatellites are short tandemly repeated nucleotide patterns occurring at very high frequencies in eukaryotic and bacterial genomes. Although short tandem repeats can be expected to occur in genomic sequences by chance, the observed number of microsatellites greatly exceeds random expectations and they attain lengths ranging from a few to thousands of repeats. It is generally accepted that the length of the repeated motif in microsatellites ranges from 1 to 6 nucleotides, whereas motifs longer than 10 nucleotides are called minisatellites. This distinction is rather artificial and thus varies among authors. For the purpose of any evolutionary discussion the distinction should be made based on mutation mechanisms and functionality rather than motif size. **See also**: Minisatellites; Repetitive DNA: Evolution

The repetitive structure of microsatellites makes them prone to errors during deoxyribonucleic acid (DNA) replication, in sharp contrast with the high fidelity required for the replication and transcription of DNA encoding for protein structures. This error propensity is termed 'microsatellite instability' and is generally attributed to the tendency of tandem repeats to undergo mutations involving insertion and deletion of whole repetitive units, at a rate that is more than 10^5 times than observed for point mutations. Microsatellite alleles are distinguished by length in base pairs (bp) and the total variation is expressed by the number of alleles generated by each locus. The magnitude of change (i.e. number of repeat units involved in a mutation) varies, but most frequently a single repeat is added or lost per mutational event. **See also**: Microsatellite Instability; Simple Repeats

Although exceptional polymorphism is the most conspicuous characteristic of microsatellites, not all microsatellites are polymorphic at the same point in time or at the same rate. This leads to the question: 'What influences microsatellite mutability?' Microsatellite mutations can occur during chromosomal replication, either as a part of mitotic or meiotic processes, or during repair or recombination processes that require DNA synthesis. Therefore the frequency of microsatellite mutations increases in rapidly dividing cells or under stress conditions when cells undergo active repair due to damage. Any mutation arising during meiosis or in the initial mitotic divisions of an embryo will proliferate and, if not selected against during development, will constitute a new microsatellite allele. In contrast, mutations arising in differentiated cells constitute somatic mutations which will not affect other cells unless the affected cell starts dividing again, as is the case in cancerous cells. To our current knowledge, many factors interact to affect microsatellite mutation rate during these processes, which have been the focus of numerous studies, but only a few have generalizable effects (**Figure 1**). **See also**: DNA Replication

The study of microsatellite mutation patterns and evolution can be intricate. On average 10^{-3}–10^{-5}

ELS subject area: Evolution and Diversity of Life

How to cite:
Vargas Jentzsch, Iris M; Bagshaw, Andrew; Buschiazzo, Emmanuel; Merkel, Angelika; and, Gemmell, Neil J (April 2008) Evolution of Microsatellite DNA. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020847

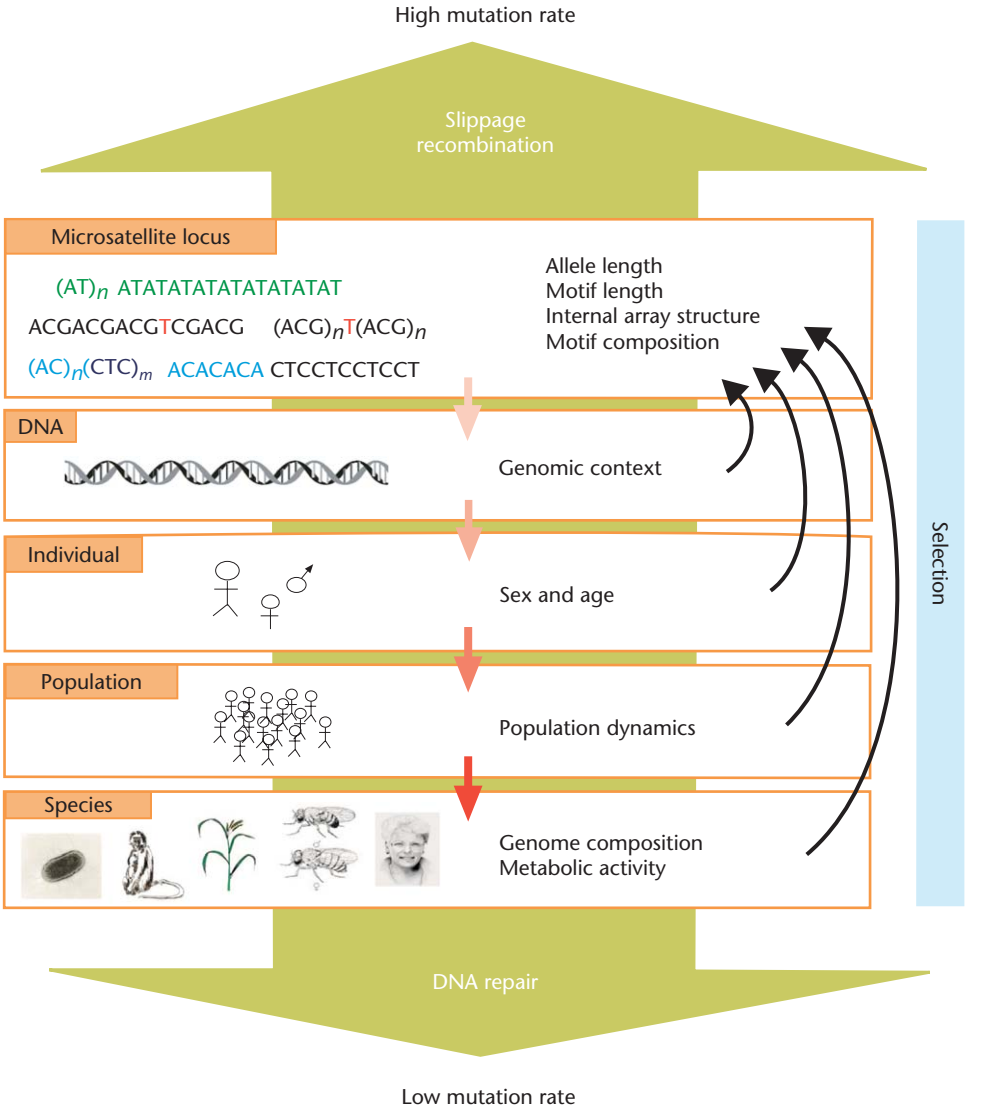


Figure 1 Factors and processes affecting microsatellite mutation. Factors (*italics*) operate at different hierarchical levels (orange boxes), starting from the smallest scale the microsatellite locus itself and moving up to the species level. Selection operates across all levels. All these factors interact dynamically, affecting the rate of replication slippage and recombination and, therefore, microsatellite variability.

microsatellite mutations arise per locus per generation, which means that, for any given locus, between one in 10^3 to one in 10^5 dividing cells will have a different microsatellite allele to that of the surrounding cellular community. The identification and quantification of microsatellite mutations is readily achieved for rapidly reproducing unicellular life forms, like bacteria or yeast, by direct screening of large numbers of individuals for microsatellite mutations. However, large-scale screening is generally not practical for complex organisms with long life cycles, and alternative strategies are adopted. A number of direct and indirect methods to study microsatellite evolution have been developed. The advantages and limitations of these study strategies are shown in **Table 1**. Additionally, the mathematical models developed to date to explain microsatellite dynamics are depicted in **Box 1**.

Microsatellite Abundance and Distribution

Microsatellites are abundant in all eukaryotic and bacterial genomes studied so far, but the validity of this statement is strongly dependent on the definition of microsatellite in terms of minimum repeat number. Microsatellites with less than 8 repeats are overrepresented in all genomic sequences, but if we increased the size threshold (10, 12, 14 repeats), bacteria and the archaeans would slip out of the

Table 1 Methods to study microsatellite evolution: usefulness and limitations

Method	Principle	Utility	Limitations
Direct methods	Based on direct observations of microsatellite mutations obtained by screening multiple generations of cells or individuals using genotyping or sequencing techniques	Provides a detailed picture of the short-term evolution of microsatellites	Requires highly mutable loci to generate a significant number of mutational events ($n > 100$)
Pedigree analysis	Analysis of allele transmissions in parent–offspring pairs, ideally across multiple generations	Direct observations of spontaneous germline mutations	Requires large pedigrees and extensive genotyping; not tailor-made to address specific questions, rather a byproduct of another large scale study
Sperm typing	Identification of germline mutations by genotyping sperm cells which are easy to obtain in quantities large enough to observe multiple mutations	Estimation of mutation rates with high confidence, thanks to the large sample sizes	Ideally, the microsatellites should be sequenced to obtain the most information out of the experiment. But standard sequencing techniques are not efficient enough. New genomic sequencing technologies could overcome this problem (e.g. 454 and Solexa)
Reporter gene systems	*In vitro* observation of loss or recovery of function of a monitored gene containing an inserted microsatellite. A succeeding frameshift would destroy the function and a subsequent one could restore it again	Rapid identification of mutations in tailor-made repeat sequences	Artificial DNA fragments do not provide a true picture of the organism sequence variation; limited by ability to culture organisms/cells of interest
Mutation accumulation lines	Study of microsatellites in immortal cell lines and tumour cells, which show abnormally elevated rates of mutation	Observation of microsatellite mutations *in vitro*, to measure mitotic rates of change; especially appropriate when comparing different cell types	Results from these systems are not directly comparable to the estimates of the germline mutation rate which is linked to meiosis
MMR-deficient systems	Use of knock-out cell lines and organisms to detect substantial changes in the rate and pattern of mutations (bacteria, yeast and mice)	Study of the role of different MMR proteins in maintaining microsatellite stability	Limited to successful knock-out systems; difficulty to compare results between systems
PCR	Use of the intrinsic instability of the *Taq* polymerase during microsatellite replication *in vitro* to study replication slippage mechanism	Provided evidence that the out-of-register alignment of strands is an intrinsic property of tandem repeats	Since it is based on *in vitro* reactions, it lacks the factors affecting the same reactions *in vivo* (e.g. MMR)
Indirect methods	Analysis of large datasets with bioinformatics tools to infer microsatellite mutations and evolutionary patterns	Allow to test available models of evolution against empirical data	Rely on demographic and evolutionary assumptions
Population data	Genotyping of unrelated individuals within a population at numerous microsatellite loci. Mutation	Testing of evolution models using computer simulations to compare the observed distribution of allele	Rests on the assumption of mutation drift equilibrium, a condition that is not necessarily met in natural

(Continued)

Table 1 Continued

Method	Principle	Utility	Limitations
	rates are then inferred from allele frequency data	frequencies to the expected distribution	populations but holds for human populations
Genomic data analysis	Inference of microsatellite evolution from unbiased genomic distributions and/or multiple sequence alignments	Allows intergenomic comparison of microsatellite distributions, motif preference, nucleotide composition, etc., to study species-specific factors affecting microsatellites. No ascertainment bias	Generally only one individual from each species has its genome sequenced, which does not yield data about polymorphism
Phylogenetic framework	Use of a phylogenetic framework to compare the sequence evolution of orthologous microsatellite loci in related species	Microsatellite allele states are characterized by sequencing and superimposed on the edges of a phylogenetic tree, and the order and character of past individual mutation events are inferred	The likelihood of finding orthologous microsatellites decreases with increasing sequence divergence and evolutionary distance between species; assignment of ancestral states is a difficult task

Box 1 Mutation models

Why Models?

All genetic markers used to assess genetic distance (e.g. in population genetics, phylogeography and phylogenetics) depend on the knowledge of the mutation processes that generate their variation, and on the robustness of the underlying estimates of mutation model parameters, such as the mutation rate or directionality. A wide range of models have been proposed to explain the mutational dynamics of microsatellites. However, this variation indicates that an integral body of theory is missing to interpret all of the available facts. This practical dilemma is illustrated by many studies where more than one option has been considered.

Infinite Allele Model

The simple infinite allele model (IAM) assumes that each mutation creates a new allele in the population. However, the forward–backward mutation process at microsatellite loci ultimately results in the creation of alleles identical in state, a condition referred to as size homoplasy. Only the unusual dynamics of compound/complex microsatellites seem to be described best by the IAM.

Stepwise Mutation Model

Under the stepwise mutation model (SMM), mutations accrue via the addition or deletion of a single repeat at a time. Gains and losses occur at equal frequency and at a rate independent of allele size. Various estimators of genetic distance based on the SMM have been developed for phylogenetic and demographic applications. Even though the SMM is adequate when closely related populations are considered, this simplistic model may be inadequate when a critical level of divergence is reached.

Two-Phase Model

The two-phase model (TPM) is an extension of the SMM that allows for infrequent multistep mutations: the one-step mutations are more likely to occur and follow the SMM, whereas the magnitude of multistep mutations follows a truncated geometric distribution. Some contention has been raised around studies that found better fits with the TPM than with the SMM, as they used allele size scored from polymerase chain reaction (PCR) product length, and thus could not account for length change mutations in the flanking regions.

Biased Mutational Process Models

A number of sophisticated models have been proposed to explain the many complexities of microsatellite mutational dynamics, e.g. dependence of the mutation rate on allele length and on the number of point mutations, mode and tempo of expansion and contraction events, directional bias and upper length constraint. However, these models have not been routinely applied to empirical data.

It is arguable whether there is one possible best model to explain variation at microsatellite loci. A fair question to ask is whether the choice of model really matters, as biologists might feel that the resolving power of microsatellites outweighs the alleged simplicity of the SMM and the TPM. But for most, the oversimplicity of assumptions contained in present theories cannot be ignored when estimating genetic distance, especially when high divergence is envisaged. The incorporation of most of the known features of microsatellite dynamics into one or more model of evolution is a first but important step towards the challenging development of an integrative and realistic theory. However, it is still unclear how much complexity can be ignored while trying to closely reflect empirical observations.

picture because longer repeats become scarce. Bacterial genomes are believed to be restricted to small sizes (<5 Mb) by strong selection for rapid replication; therefore repeats would be removed unless they are favoured by selection. However, as we discuss later, there are numerous examples of microsatellites in bacteria favoured by selection to be long and therefore polymorphic, which serve as sources of functional diversity within coding regions.

Comparative analyses of microsatellite abundance among genomes can also be biased due to conceptual issues. Microsatellites are usually regarded as perfect tandem repetitions of a given motif. However, the majority of microsatellites include 'imperfections' in the form of insertions, deletions or substitutions. The accumulation of imperfections frequently leads to multiple stretches of perfect tandem repetitions imbedded within longer imperfect versions of the same motif, or microsatellites composed by more than one motif (composite microsatellites). For this section and any discussion of microsatellite evolution we refer to perfect and imperfect motifs as the extremes of a continuum.

The total content of microsatellites (longer than 15 bp) varies among genomes, ranging in mammals from 2% (horse) to more than 5% in dog, mouse and rat. Closely related genomes like the human and chimpanzee possess very similar microsatellite contents (3.44 and 3.47% of the genome, respectively) and, as taxa diverge, microsatellite content diverges too. This divergence is clearly not a function of total genome size which, especially in higher eukaryotes, varies as a function of interspersed repeat content. Rather microsatellite content seems to be associated with conserved and unique sequences (Morgante *et al.*, 2002). **See also**: Chromosomes: Noncoding DNA (Including Satellite DNA)

Within a single genome, the total density of microsatellites is usually very similar among chromosomes, although sex chromosomes and smaller chromosomes in eukaryotes appear frequently as outliers, showing higher microsatellite content relative to the rest of the chromosomes. The distribution of microsatellites within each chromosome is also very homogeneous. With the exception of centromeric, subtelomeric regions and telomeres, microsatellites occur at very constant intervals, the interval size depending on the total density of microsatellites (e.g. approximately 0.85–1 microsatellite per kilobase (kb) in human chromosomes when taking a minimum length of 15 bp). Centromeres and subtelomeric regions are composed of repetitive sequences with longer motifs, and devoid of microsatellites; subtelomeric regions are almost completely covered by minisatellites. Telomeres, sequences at both ends of linear chromosomes, are completely covered by microsatellites. In association with specialized proteins they prevent chromosome shortening due to incomplete DNA replication at these ends. The repeated motif in vertebrate telomeres is TTAGGG (T, thymine; A, adenine; G, guanine) and the total length of these repeats varies among species and depends on cell type; ranging from 5 to 15 kb in humans, and up to 200 kb in laboratory mice. These lengths are approximate since they cannot be determined precisely due to the difficulty in sequencing repetitive DNA. **See also**: Telomeric and Subtelomeric Repeat Sequences

Differences in distribution among genomic regions (introns, exons, intergenic) are observed when classifying microsatellites by motif length and nucleotide composition. Tri- and hexanucleotide microsatellites are overrepresented in coding regions in comparison to their representation in introns and intergenic regions. Coding sequences need to be transcribed accurately to conserve the reading frame of the encoded proteins. Since the genetic code is made up of triplets any nucleotide insertion or deletion which is not a multiple of three would disrupt the reading frame. Furthermore, not all combinations of triplet repeats are present in coding regions. The motifs of the most common repeats in mammals translate into amino acids with a mixture of characteristics: polar amino acids such as glutamine (most commonly encoded by CAG; C, cytosine), serine (AGC) and glutamic acid (GAG); and nonpolar amino acids-like proline (CCG), leucine (CTG), glycine (GGC) and alanine (GCG). Glutamine repeats are exceptional in that they very often expand beyond 20 repeats, and in these cases they are frequently involved in the development of neurodegenerative diseases (more than 100 repeats). In contrast, triplets with high content of T are scarce or absent; none of the 10 codons containing more than one T in the sense strand (the strand that gets transcribed) is reiterated. Finally triplet repeats with motifs ACT and ATC are absent because they translate into stop codons. Therefore, microsatellite motif abundance within coding regions depends on codon usage and selective constraints on proteins, which can be different among species.

Microsatellites within exons tend to be more C/G rich, as coding regions in general have higher C + G content. However, introns and intergenic regions contain in general more mono-, di- and tetranucleotide microsatellites, with a predominance of A/T-rich motifs regardless of species. Dinucleotide repeats are the most abundant motif after mononucleotides and show the most conspicuous difference among life kingdoms: the AC motif predominates in mammalian genomes in contrast to AT, which is clearly the most abundant motif in plants and most fungi.

Mechanisms of Microsatellite Mutation

The mutational mechanisms of microsatellite array length change have been studied extensively at a molecular level. Two types of model have traditionally been invoked, both of which involve the ability of DNA strands with tandem repeat sequences to stably align out of register with respect to flanking sequence. Early theorists (Smith, 1976) believed that microsatellite evolution could be driven by unequal recombination during meiotic cell division through misalignment of homologous chromosomes (**Figure 2**, Part A).

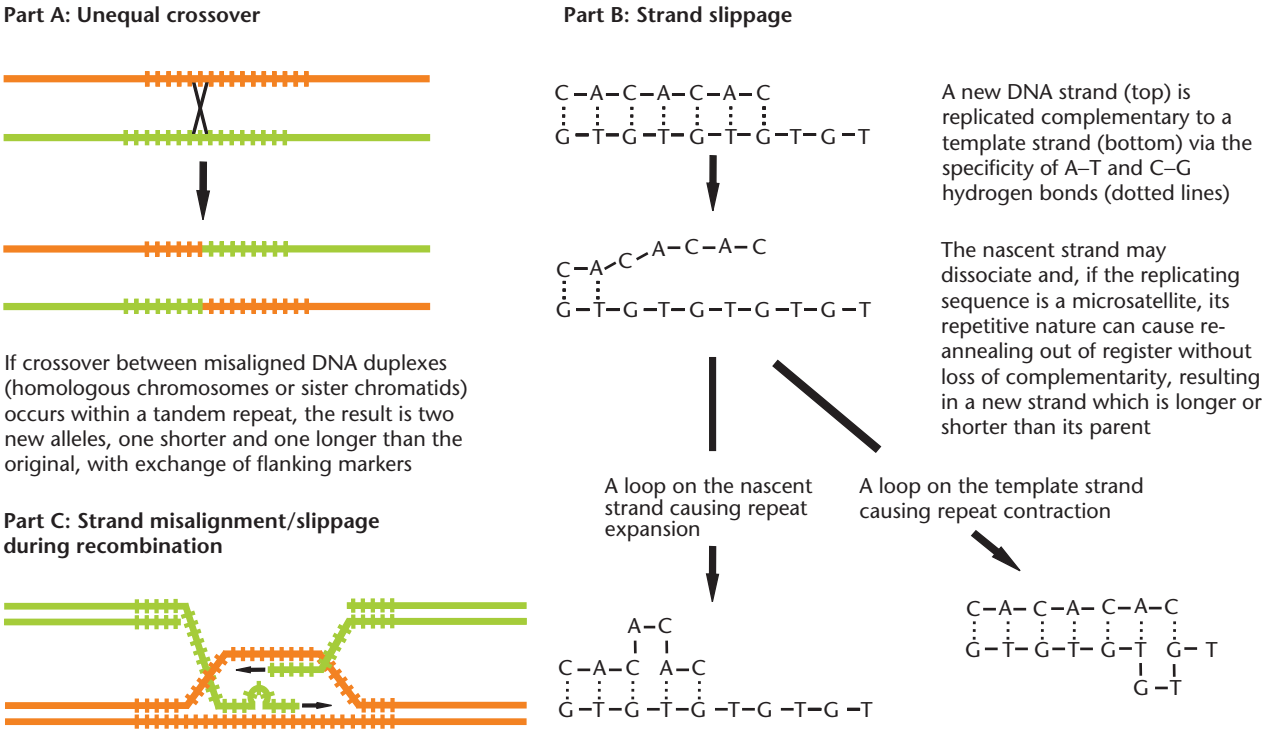


Figure 2 Models of tandem repeat length mutation. Unequal crossover, involving misalignment of homologous chromosomes or sister chromatids (Part A) and strand slippage (Part B) are the two main types of mechanism that have been proposed. Strand slippage can occur during any process requiring DNA synthesis, including recombination (Part C).

More recently, evidence has accumulated that misalignment of single strands during DNA replication is a more common mechanism of microsatellite mutation (**Figure 2**, Part B). Other data indicate, however, that recombination does have a role in some instances.

Slipped strand mispairing, or replication slippage, was first proposed by Levinson and Gutman (1987) who found that *Escherichia coli* mutants deficient for recombination showed no reduction in microsatellite instability. Similar results were later found in yeast. According to the replication slippage model, recombination is not required to explain microsatellite mutations. The misalignment that gives rise to mutations is between a newly synthesized DNA strand and its complementary template strand. This can occur if the two strands dissociate (slip) and then reanneal out of register, forming a loop, which is stable because, due to the repetitive nature of the sequence, it can be formed without deviation from normal Watson–Crick base pairs. If the loop is formed on the nascent strand, the resulting mutation will be a repeat expansion, while loops in the template strand cause reduction in repeat length. Such loops can be repaired by the well-characterized mismatch repair (MMR) system and it has been shown that MMR deficiency is associated with microsatellite instability in a variety of organisms. Additional support for a role for replication errors in microsatellite mutations is seen in *in vitro* replication of DNA strands containing microsatellites. This often results in a proportion of replicated tracts with higher and lower repeated copy numbers than the original.

Other lines of evidence also suggest that replication errors play a greater role in microsatellite evolution than unequal recombination. Microsatellite mutation rates are significantly higher than recombination crossover rates, and are similar on the nonrecombining human Y-chromosome and the recombining autosomes. Interpretation of these results is not straightforward, however, because of high rates of recombination events that do not involve crossing over between chromosomes, which are difficult to detect, and have been shown to occur on the Y-chromosome as well as on the autosomes. Exchange of flanking markers, which would be expected in unequal crossover-associated mutations (**Figure 2** Part A), has generally not been seen for microsatellite length mutations, unlike for minisatellite and satellite repeats, which have repeat motifs between 10 and several hundred base pairs. This may be due to the fact that minisatellites and satellites are often many fold longer than microsatellites, allowing more opportunity for homologous chromosomes or sister chromatids to misalign during meiosis. Evidence has, however, implicated unequal recombination without exchange of flanking markers (gene conversion) in the mutation of some microsatellites involved in human disease. Furthermore, microsatellite abundance is positively correlated with recombination rate at scales ranging from kilobases to megabases, though some evidence suggests that this may be

attributable to stimulation of recombination by microsatellites. **See also**: Eukaryotic Recombination: Initiation by Double-strand Breaks; Meiotic Recombination Pathways

The question of whether heterogeneous mutational mechanisms are a significant factor underpinning the differences in mutation rates and patterns between microsatellite loci remains open to some degree. Most of the studies looking closely at mutational mechanisms have been of those microsatellites involved in heritable human disease or cancer, and these may represent a biased sample (see **Box 1**). Further experimental studies will be necessary to quantify the relative influence of recombination and slipped strand mispairing on microsatellite mutation. The distinction is complicated by links between the two processes, since strand misalignment can occur during the stages of recombination that involve strand exchange between chromosomes and subsequent synthesis of DNA (**Figure 2**, Part C). One line of inquiry that may prove to be important is the possible influence of recombination hotspots, as microsatellite mutability in these narrow, frequently recombining, and evolutionarily labile regions has not been extensively sampled.

Factors Influencing Microsatellite Mutation Rate

The rate at which microsatellites undergo mutations can be affected by multiple factors influencing either the probability of generation of mutations (by replication slippage or recombination) or the efficiency of repair of these mutations. Mutation and repair are constant antagonistic processes, which themselves are constrained by selection. Most microsatellite mutations arising by replication slippage are quickly corrected, mainly by the MMR system, rendering MMR efficiency as one of the main factors affecting microsatellite mutation rate. The efficiency of repair declines as the size and/or stability of the loop formed during replication slippage increase. These two characteristics, in turn, are affected by other factors intrinsic to the microsatellite: the allele length, motif length, nucleotide composition, imperfections within the microsatellite and the genomic position of the repeat. The probability of recombination events within microsatellites is also affected by these factors. Furthermore, the interactions among these factors are dynamic and take place at different levels simultaneously, as depicted in **Figure 1**. The probability of transmission of microsatellite mutations to offspring also depends on individual sex and age. Additionally, the fact that the distribution and abundance of microsatellites differs among species suggests that all these factors interact with species-specific metabolic characteristics.

Array length

The most accepted factor influencing microsatellite mutation is overall array length. The conventional wisdom is that longer arrays have higher probability for misalignment during replication or recombination, therefore longer repeats tend to have higher mutation rates. Additionally, the direction of mutation varies with array length; i.e. short microsatellites tend to experience more expansions than contractions, whereas longer microsatellites are occasionally subject to large deletions. The first case, where microsatellites tend to expand, could be explained by a bias of loop formation in the leading strand during DNA replication. Large deletions, however, are more likely to occur by recombination; the longer the microsatellite, the higher the probability of nonhomologous alignment during meiotic recombination. However, the 'directionality' of microsatellite mutation is not clear yet, and could be produced by the interaction of several other factors.

Internal array structure

Point mutations and other interruptions within the repeat array have been observed to reduce mutation rate, which is most likely due to an overall reduced chance of slippage, secondary structure formation and/or recombination. Repeat arrays with several motifs, known as compound or complex/clustered microsatellites, show elevated mutation rates for at least one of the internal motifs, when compared to a microsatellite of the same length and motif. One explanation for this is that both fractions, although containing different motifs, have similar structural propensities.

Motif nucleotide composition

Repeats with certain motifs have a heightened propensity to form secondary structures to alter DNA structure. Secondary structures, such as hairpins, quadruplex structures, H-DNA or sticky DNA, being intermediate DNA hybrid forms, increase the likelihood of strand misalignment and subsequent polymerase slippage, whereas a conformational change in DNA structure, such as Z-DNA formed by long AC tracts, will affect polymerases and repair enzymes alike. Some sequences, like the spinocerebellar ataxia-causing $(ATTCT)_n$ element, even have the potential to unwind DNA locally, promoting single-stranded DNA which highly facilitates secondary structure formation. Nevertheless, microsatellites with structure-forming potential are often not the most abundant array types across genomes and sometimes show the slowest *in vitro* slippage rates.

Motif length

Shorter microsatellite motifs allow for more opportunities of misalignment than longer motifs. Motifs more than three nucleotides long require higher dissociation energy, and are thus less likely to generate enough single-stranded DNA to form a stable loop. Furthermore, motif length can affect MMR efficiency. If the loop is too big (i.e. more than 18 bp), the efficiency of MMR drops. Here, the effect of motif length becomes evident, since longer motifs form bigger loops.

Genomic context

The mutability of any DNA sequence depends on its context within a genomic sequence. This is most apparent when observing the distribution of microsatellites in coding regions where the effect of mutations has a high probability of being disadvantageous and is therefore strongly counteracted by selection. Alternatively, mutation rate variation can arise through structural propensities of either flanking sequences or even more distantly neighbouring regions, being most likely based on the thermodynamic propensities of different base-pairings ($\Delta T_m GC > \Delta T_m AT$). A few studies have shown that the propensity of expansion of certain types of microsatellites, namely GC-rich trinucleotide repeats, is positively correlated with GC-bias of the flanking sequence, but others found no evidence for such a correlation. Further, CpG islands (CG dinucleotides) are found in many mammalian promoters and are, when methylated, involved in chromatin remodelling and gene silencing. The observed proximity of some highly expandable loci to CpG islands has led to the suggestion of a mechanistic link between these elements and microsatellite instability.

Sex and age

Sex and age affect the probability of transmission of a mutated allele. For example, in humans, males produce considerably more gametes than females throughout their lifetime, increasing therefore the cumulative germline cell divisions and associated mutations with age. In contrast, the female reproductive system stops producing ovules after birth, therefore being exposed to fewer mutations associated with DNA replication, and having no significant age effect. Supporting studies have found that male reproductive cells mutate five times more often than female ones, and older men pass on more mutations than younger men. However, studies in other species, for example fish, show fewer differences among male and female transmitted mutations, because the ratio of male to female gametes is smaller.

DNA repair

DNA repair is essential in maintaining DNA integrity and to prevent mutations. Failure of the MMR system during replication results in up to 10^3-fold increase in microsatellite instability. Microsatellite instability was the first clue indicating a failure in MMR in certain types of tumours, like colorectal cancer. Defects in the exonucleolytic proofreading activity of DNA polymerases have less impact on microsatellite mutation, with a 5–10 fold increase in mutation rate. Further, repair activity is not uniform throughout the genome, e.g. highly transcribed genes experience stricter repair than others. MMR has been found to be strand and substrate (sequence) specific, but it is not yet clear to what extent this specificity affects microsatellite mutation. **See also**: DNA Repair

Origin of Microsatellites

One of the main hypotheses proposed to explain microsatellite genesis regards the fortuitous generation of repeated motifs within random sequences by point mutations or small insertions and deletions. Once a 'proto-microsatellite', with two or three repeats, has arisen, its maintenance and growth is expected to be favoured by its propensity to undergo strand slippage during replication and, depending primarily on the repeat motif, its capacity to form unusual DNA conformations and to participate in recombination and transposition events. As discussed earlier, the number of repeat units correlates positively with the mutability of the microsatellite, but the minimum repeat number needed to allow for strand slippage or other mechanisms involved in microsatellite mutation to occur is debatable. Initially, eight repeats were suggested as the minimum threshold for a small tandem repeat to be considered a microsatellite, and therefore smaller microsatellites were left out of most studies in eukaryotes. In bacteria, microsatellites with less than eight repeats were shown to undergo appreciable rates of mutation, and microsatellites with as few as two repeats were shown to be polymorphic in *Mycobacterium* species (Sreenu *et al*., 2006).

A second widely accepted hypothesis regards the dispersion of sites for microsatellite origin by transposable elements (especially retrotransposons). Transposable elements are sequences that have the capacity to 'jump' (transpose) to different positions in the genome generating multiple copies of themselves. These can be divided into two main classes based on their mechanisms of movement. Class I are retrovirus-like transposons that get transcribed into messenger ribonucleic acid (mRNA) and subsequently retro-transcribed back to DNA and inserted in a new position in the genome. Class II are so called cut and paste transposons because they get excised from their original position and inserted into a new position. Both of these elements can leave traces of their presence and movement during the transposition process across DNA sequences, which resemble microsatellites (e.g. contain small tandem repeats), especially poly A arrays. Class I retrotransposons get a poly A tail added at the 3′ end after transcription into mRNA, which then gets inserted together with the transposed sequence into the new position. Retrotransposons can also contain other microsatellite-like stretches within their sequences including dinucleotide and tetranucleotide repeats. Class II transposons insert preferentially into certain DNA sequences which can be either inverted repeats or tandem repeat sequences. This suggests a reciprocal association in which microsatellites act as 'retroposition navigator sequences' while retrotransposons generate more microsatellites during their dispersion throughout the genome. **See also**: Transposons as Natural and Experimental Mutagens

A good example of retrotransposon mediated microsatellite genesis in humans is the well-documented origin of A/T-rich microsatellites with motifs ranging from one to six nucleotides in length from *Alu* elements. *Alu* repeats are the most abundant interspersed repetitive elements in primate

genomes, and are comprised of two monomers separated by a poly A tract. These retrotransposons also have the typical poly A tail at their 3′ end. Both of these repeats give rise to poly A and A-rich microsatellites (i.e. AAC, AAG, AAAAT), and dinucleotide microsatellites (i.e. AT, AC, AG). The 3′ end poly A tail tends to be longer than the middle one in humans, giving rise to the major part of microsatellites arisen from *Alu* elements.

The association of poly A and A/T-rich microsatellites with transposable elements may, at least partly, explain the fact that A/T rich motifs are by far the most abundant repetitive arrays within genomes. In contrast, GC-rich microsatellites, especially trinucleotide and hexanucleotides do not seem to be associated with transposable elements. Rather it was suggested that the origin of trinucleotide repeats could be associated with the process of codon reiteration in the evolution of proteins, a process which favours increases in protein size by expansion of repetitive domains.

Functional Microsatellites and Evolution

For a great part of the last half a century the functional genome was regarded as those sequences coding for proteins and describing high conservation (i.e. high sequence similarity) among taxa. Thus functional analyses were focused on coding regions that, at least in human, mouse and chimp, account for less than 2% of the genome. The fact that coding exons usually lack microsatellites, or contain mostly trinucleotide repeats, led to the idea that microsatellite variation is either deleterious or restricted to nonfunctional intergenic DNA.

The adaptive value of microsatellite polymorphism was first explored in detail in pathogenic bacteria (Moxon *et al*., 1994). Genetic variation in pathogenic bacteria is generally restricted because infections often result from propagation of a few founder cells. Low genetic and phenotypic variability would render the pathogens highly vulnerable to changeable immune responses in the host. Therefore, pathogenic bacteria need to evolve at an accelerated rate. A small amount of hypermutable genes confer them this essential ability, thanks to the presence of microsatellites within their coding or regulatory regions. Since microsatellites mutate frequently and reversibly, they can inactivate the functional domain of a gene (i.e. by frameshift mutation) in one generation and mutate back to reactivate it in the next one. Within regulatory regions the effect is indirect by inhibiting or enhancing the transcription of a gene. Many of these genes code for surface molecules, called 'virulence factor genes', which have an essential role in the cell's interactions with its environment, and therefore a very high impact on microbial fitness. Examples of virulence factor genes are genes encoding for proteins in the capsule which confers serum resistance, pili proteins which affect cell adhesion, and other surface proteins affecting the formation of surface pores and nutrient acquisition. In *Haemophilus influenzae* and *Staphylococcus aureus*, changes in the length of microsatellites within virulence factor genes result in conformational changes in processed surface proteins, which make these unrecognizable to the host's antibodies. A similar situation can be observed in eukaryotes such as the yeast *Saccharomyces cerevisiae*, in which more than 75% of genes containing microsatellites encode cell wall proteins. Variation in repeats associated with these genes gives rise to quantitative alterations in phenotypes (e.g. adhesion, flocculation or biofilm formation) (Verstrepen *et al*., 2005). **See also**: Mutagenesis Mechanisms

In populations of single-celled species, mutations generating variability can immediately favour adaptation by increasing the chance that at least a few individuals will be able to survive under stress conditions. If a microsatellite mutation is deleterious, the death of a single cell will not endanger the rest of the population. However, multicellular organisms should be more intolerant to mutations in microsatellites or elsewhere because each cell is part of a complex system. This problem was partly overcome by increasing proteome diversity by segmental duplications, diploidy and polyploidy, and the implementation of alternative splicing, which requires a concomitant increase in regulatory information. In contrast to single celled organisms, multicellular eukaryotes have extensive intronic and intergenic sequences whose extent increases with developmental complexity (Taft *et al*., 2007). Interestingly, the majority of the genome (including noncoding regions) gets transcribed during some stage in development, but most of these transcripts, including a great proportion of transcripts from coding regions, are not translated into proteins. Instead, they constitute introns, 5′ and 3′ UTRs (untranslated regions), or remain as RNA regulating cell functions. **See also**: Chromosomes: Noncoding DNA (Including Satellite DNA); Transcriptional Regulation: Evolution

Owing to the intricate ways in which DNA sequences and protein complexes interact to fulfil cellular functions, the repetitive structure and frequent length variation of microsatellite sequences, both within and outside coding regions, can influence chromosome structure, gene expression, protein function and even DNA repair and recombination in multiple ways which are broadly outlined in **Figure 3**, and described later.

Effects of microsatellites within exons

Coding repeats, especially microsatellites, are targeted by numerous studies because their instability can lead to genetic diseases such as Huntington disease and Friedreich's ataxia. However, the relatively high incidence of trinucleotide microsatellites within exons suggests that these are not being eliminated by selection due to potential benefit for the cell. Indeed, microsatellites are markedly overrepresented in transcription factors, protein kinases and genes encoding developmental regulatory proteins.

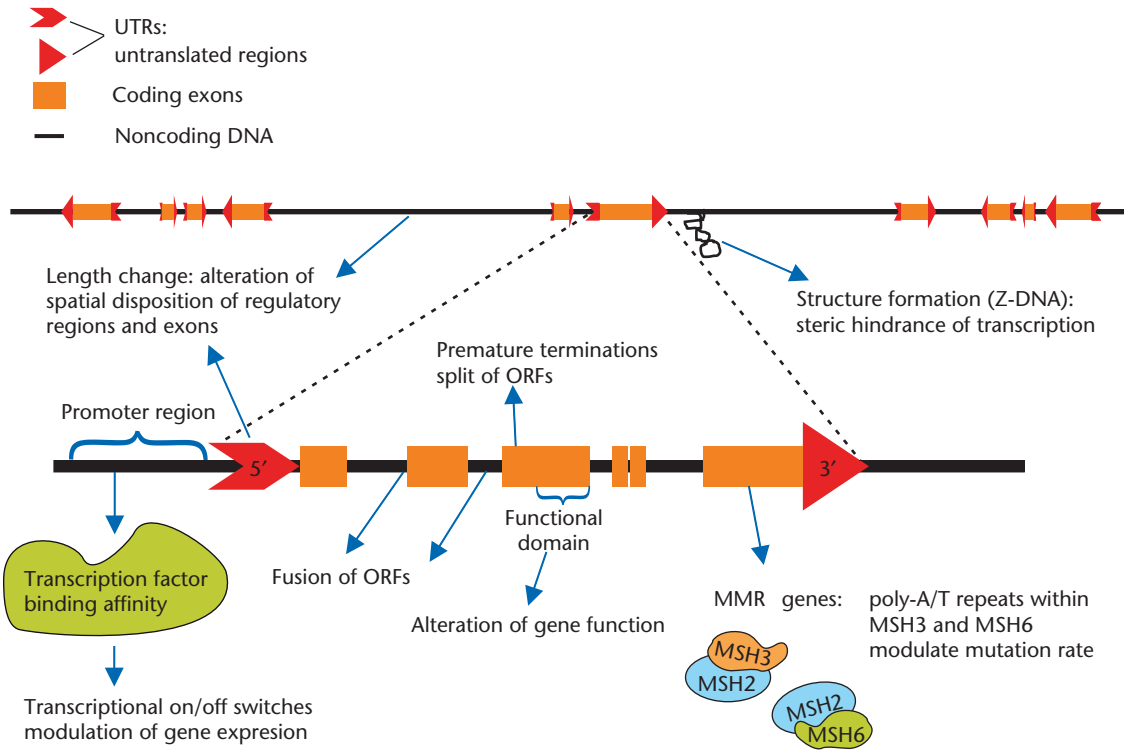


Figure 3 Functional implications of microsatellite length change. Microsatellite length variations have been shown to mediate diverse functions depending on the genomic region in which these are present. Within exons microsatellite mutations can induce changes in protein structure, therefore altering its function, or can directly inactivate the protein by trunctation or fusion of open reading frames (ORFs). Within introns and intergenic regions these changes can partake in the modulation of gene expression, either by modifying the structure of transcription factors or enzymes involved in transcription modulation, or by changing the secondary and/or tertiary structure of DNA or RNA regions that interact with transcription factors. Furthermore, microsatellites are involved in the regulation of their own and genome-wide mutation rates, by being present within the minor components of the mismatch repair system.

Fondon and Garner (2004) hypothesized that the impressive range of phenotypic variation observed among dog breeds was due to length variations of microsatellites within developmental genes. As part of their study they found that *Runx-2*, a gene coding for a transcription factor which, in vertebrates, regulates the differentiation of osteoblasts, has two homopolymeric tracts side by side within its amino acid sequence: polyglutamine (18–20 repeats) and polyalanine (12–17 repeats). Coded by two perfect trinucleotide repeats, the ratio of repeat lengths of these alleles correlated strongly with the downward bending of the dog's muzzle in several dog breeds. Another gene, the *Alx-4* gene, contains an imperfect hexanucleotide repeat coding for poly proline–glycine stretch. A 51-bp deletion in this repeat destroys the binding ability of the *Alx-4* protein to bind to the lymphoid enhancer binding factor-1 to target gene expression in limb bud mesenchyme (Fondon and Garner, 2004). The consequence in both mice and dogs is the development of an additional digit in the hind feet (polydactyly). Like these two examples, an increasing number of studies are uncovering the effects, either positive or negative, of repeat expansions within exonic regions. Polyglutamine peptides have been shown to drive transcription while polyalanines repress transcription in a length dependent fashion. The processes affected are generally regulatory, influencing directly or indirectly the expression of proteins which act at different levels of enzymatic cascades.

Another example, where the effect of microsatellite mutation within genes has genome-wide effects, is the case of microsatellites within the mismatch repair system genes. These encode an enzymatic complex which is highly conserved with close homologues between eukaryotes and bacteria and archaeans, and is involved in the correction of base pair mismatches and mutations due to strand slippage and loop formation during replication. The coding regions of the minor MMR genes (hMSH3 and hMSH6 in humans) contain several mononucleotide repeats (mainly poly A/T, adenine/guanine), and variations in the length of these stretches permit the modulation of mutation rates over evolutionary time. The MMR system is normally extremely efficient and, therefore, microsatellite length in somatic cells tends to be stable. However, if the MMR system becomes defective or overwhelmed due to external factors (e.g. mutagenic agents), cells start accumulating altered microsatellites by thousands, a phenomena known as microsatellite instability, which is involved in cancer development. **See also**: Mismatch Repair Genes

Effects of microsatellites in introns and noncoding regions

Noncoding DNA might contain the majority of regulatory DNA. The concept of regulatory region is not yet well defined; a promoter region, for example, is a site in DNA where RNA polymerase binds to start transcription. The promoter could be several kilobases away from the transcription start site and is generally difficult to recognize based on DNA sequence only. Furthermore, regulatory sequences seem to have an elevated turnover rate; transcription factor-DNA interactions are highly polymorphic, and regulatory interactions are constantly gained and lost within populations. On average, humans are heterozygous at more functional *cis*-regulatory sites ($>16\,000$) than at amino acid positions ($<13\,000$), in part because of an overrepresentation among the former in multiallelic tandem repeat variation, especially AC (adenine–cytosine) dinucleotide microsatellites. The role of microsatellites in gene expression variation may provide a larger store of heritable phenotypic variation, and a more rapid mutational input of such variation than has been realized (Rockman and Wray, 2002). **See also**: Transcriptional Regulation: Evolution

An interesting example is the involvement of a complex microsatellite (e.g. $(CAGA)_n$, $(CATA)_n$, $(AG)_n$ and $(GAGGAGA)_n$ interspersed among nonrepetitive sequences) in the modulation of social behaviour. The microsatellite is immersed in the 5′ regulatory region of the vasopressin 1a receptor (V1aR), which mediates the expression of the hormone vasopressin. Among other functions, vasopressin is implicated in memory formation and social behaviour in vertebrate species. Varying degrees of social interaction in voles (genus *Microtus*) were found to correlate with differing levels of vasopressin receptor expression in the brains of these species, and this in turn, with the size of the microsatellite (Hammock and Young, 2005). Praire and pine voles have a long version of the microsatellite (430 bp in total), and show high levels of social interest (i.e. the males are monogamous). In contrast, montane and meadow voles, which possess a truncated version of the microsatellite, are socially indifferent and the males do not contribute to parental care. Further, the capacity of the microsatellite to drive V1aR expression was demonstrated by *in vitro* luciferase reporter assays. In humans, four polymorphic microsatellites surround the human vasopressin reporter homologue, which suggests that behavioural variation in humans is likely to be subject to complex and highly variable regulatory interactions.

Microsatellites can also affect the structural properties of DNA. Expansions or contractions of a microsatellite change the length of the DNA sequence and consequently the spatial disposition of transcription factor binding sites with respect to exons and other transcription factors. Furthermore, the structure-forming potential of tandem repeats has the capacity to generate steric effects, favouring or disfavouring the access of transcription enzymes to particular coding regions. **See also**: DNA Structure: Sequence Effects

One of the key applications of microsatellites as molecular markers is the construction of linkage maps for gene and quantitative trait loci (QTL) mapping. These applications are rooted on the major assumption of microsatellite neutrality, and microsatellite variation is used to identify linked genomic regions possibly involved in the generation of quantitative phenotypic variation. Recent evidence on microsatellite functionality, especially the potential of microsatellites to be involved in multiple processes of gene regulation, suggests the possibility that those microsatellites associated with QTLs are the actual effectors of the phenotypic variability observed in QTL analyses. **See also**: Quantitative Trait Loci (QTL) Mapping

Evolution is a trade-off between gaining diversity in function and escaping the deleterious effects of mutations. Natural selection will favour the 'fittest' individuals within a population, but which individuals are the fittest can be redefined suddenly depending on environmental influences. Because environmental changes occur stochastically and are unpredictable, fitness is dependent upon the available diversity in any limiting characteristic during situations of stress. In these situations, high mutability can be useful for the generation of genetic diversity. However, accumulation of random mutations where most of these are likely to be deleterious is likely to reduce fitness. Microsatellite mutations affecting protein function or expression can be regarded as 'strategic mutations' because, besides of occurring at higher rates, these length mutations are gradual and fully reversible, and are ubiquitously available, therefore enabling rapid evolutionary adaptation. **See also**: Mutation–Selection Balance

The majority of microsatellites across a genome might not have a defined or critical role, because microsatellite sequences are likely to arise and expand at higher rates than their recruitment for functionality. However, variation in microsatellites is generated constantly and constitutes a rich reservoir of genetic variation. It is the intrinsic variation within these sequences, both in functional and nonfunctional regions, that underlies the evolutionary importance of microsatellites.

Concluding Remarks

Despite the initial view that microsatellites are simple, random and unstable accumulations of tandem repeats, there is now plenty of evidence demonstrating the heterogeneous and complex nature of microsatellites. The image of microsatellites is transforming from junk DNA to important elements in genome function and evolution. Therefore, in the light of recent studies of microsatellite functionality, the effect of genomic location and selective pressure should be given more importance over the classic factors assumed to affect microsatellite mutability such as allele length, motif length and nucleotide composition. The implicit variation of microsatellites provides an abundant yet relatively safe

supply of raw material for rapid adaptation in both coding and noncoding regions of the genome.

References

Fondon III JW and Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the USA* **101**: 18058–18063.

Hammock EA and Young LJ (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**: 1630–1634.

Levinson G and Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**: 203–221.

Morgante M, Hanafey M and Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**: 194–200.

Moxon ER, Rainey PB, Nowak MA and Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current Biology* **4**: 24–33.

Rockman MV and Wray GA (2002) Abundant raw material for *cis*-regulatory evolution in humans. *Molecular Biology and Evolution* **19**: 1991–2004.

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* **191**: 528–535.

Sreenu VB, Kumar P, Nagaraju J and Nagarajaram HA (2006) Microsatellite polymorphism across the *M. tuberculosis* and *M. bovis* genomes: implications on genome evolution and plasticity. *BMC Genomics* **7**: 78.

Taft RJ, Pheasant M and Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* **29**: 288–299.

Verstrepen KJ, Jansen A, Lewitter F and Fink GR (2005) Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**: 986–990.

Further Reading

Buschiazzo E and Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays* **28**: 1040–1050.

Caporale LH (2006) *The implicit genome*. New York: Oxford University Press.

Kashi Y and King DG (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* **22**: 253–259.

Nikitina TV and Nazarenko SA (2004) Human microsatellites: mutation and evolution. *Russian Journal of Genetics* **40**: 1065–1079.